Annual Review of Statistics and Its Application

Calibrating the Scientific Ecosystem Through Meta-Research

Tom E. Hardwicke,[1,2] Stylianos Serghiou,[2,3] Perrine Janiaud,[2] Valentin Danchev,[2] Sophia Crüwell,[1,5] Steven N. Goodman,[2,3,4] and John P.A. Ioannidis[1,2,3,4,6]

[1]Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité–Universitätsmedizin Berlin, 10178 Berlin, Germany; email: tom.hardwicke@charite.de

[2]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California 94305, USA

[3]Department of Health Research and Policy, Stanford University, Stanford, California 94305, USA

[4]Department of Medicine, Stanford University, Stanford, California 94305, USA

[5]Department of Psychological Methods, University of Amsterdam, 1018 WS Amsterdam, Netherlands

[6]Departments of Biomedical Data Science, and of Statistics, Stanford University, California, USA

Annu. Rev. Stat. Appl. 2020. 7:7.1–7.27

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041104

Keywords

meta-research, meta-science, methodology, bias, reproducibility, open science

Abstract

While some scientists study insects, molecules, brains, or clouds, other scientists study science itself. Meta-research, or research-on-research, is a burgeoning discipline that investigates efficiency, quality, and bias in the scientific ecosystem, topics that have become especially relevant amid widespread concerns about the credibility of the scientific literature. Meta-research may help calibrate the scientific ecosystem toward higher standards by providing empirical evidence that informs the iterative generation and refinement of reform initiatives. We introduce a translational framework that involves (*a*) identifying problems, (*b*) investigating problems, (*c*) developing solutions,

and (*d*) evaluating solutions. In each of these areas, we review key meta-research endeavors and discuss several examples of prior and ongoing work. The scientific ecosystem is perpetually evolving; the discipline of meta-research presents an opportunity to use empirical evidence to guide its development and maximize its potential.

1. INTRODUCTION

Meta-research (research-on-research) is a burgeoning discipline that leverages theoretical, observational, and experimental approaches to investigate quality, bias, and efficiency as research unfolds in a complex and evolving scientific ecosystem (Ioannidis et al. 2015, Ioannidis 2018a). Meta-research is becoming especially relevant amid warnings of a transdisciplinary "credibility crisis" and concerns of suboptimal and wasteful applications of the scientific method (Altman 1994, Baker 2016, Chalmers & Glasziou 2009, Ioannidis 2005, Leamer 1983, Pashler & Wagenmakers 2012, Open Sci. Collab. 2015). These concerns have inspired reform initiatives intended to achieve higher standards of efficiency, quality, and credibility in science (Ioannidis 2014, Miguel et al. 2014, Munafò et al. 2017, Nelson et al. 2018, Poldrack et al. 2017).

1.1. Defining Meta-Research

Meta-research has been defined as "the study of research itself: its methods, reporting, reproducibility, evaluation, and incentives" (Ioannidis 2018a, p. 1). Given this definition, its boundaries are broad, and its thematic areas interact with several other disciplines. Scientific fields are fuzzy categories with flexible, porous, overlapping borders (Börner et al. 2012). Neighboring disciplines to meta-research include, but are not limited to, philosophy of science, history of science, sociology, research synthesis (e.g., meta-analysis), data science, journalology, bibliometrics, ethics, behavioral economics, and evidence-based medicine. All of these fields may to some extent share the goals of describing, evaluating, and improving the scientific ecosystem.

1.2. Historical Roots

Meta-research has deep roots in the beginnings of the scientific method, when intellectuals such as Francis Bacon argued for greater experimentation, openness, and collaboration (Sargent 1999). These early efforts to scrutinize and modify the scientific ecosystem were still largely based on philosophical considerations rather than systematic empirical research. Over the past century, concerns about research quality have repeatedly flared across scientific disciplines, including psychology (Elms 1975), economics (Leamer 1983), and biomedicine (Altman 1994). In parallel, systematic investigations of topics, such as publication bias (Sterling 1959), experimenter bias (Rosenthal 1966), and statistical power (Cohen 1962), have reflected a growing shift toward empiricism: an acknowledgment that mostly theoretical arguments about research practices, methods, and bias should eventually be confronted with empirical data (Faust & Meehl 2002). Initiatives such as the Cochrane Collaboration in the domain of evidence-based medicine have achieved some success at addressing suboptimal research quality; however, overall, reform efforts have often failed to gain traction. Nevertheless, the recent credibility crisis has sparked a transdisciplinary discussion (Chalmers & Glasziou 2009, Miguel et al. 2014, Nosek et al. 2015, Poldrack et al. 2017), prompted a cascade of reform initiatives (Ioannidis 2014, Munafò et al. 2017), and catalyzed the emergence of the meta-research discipline.

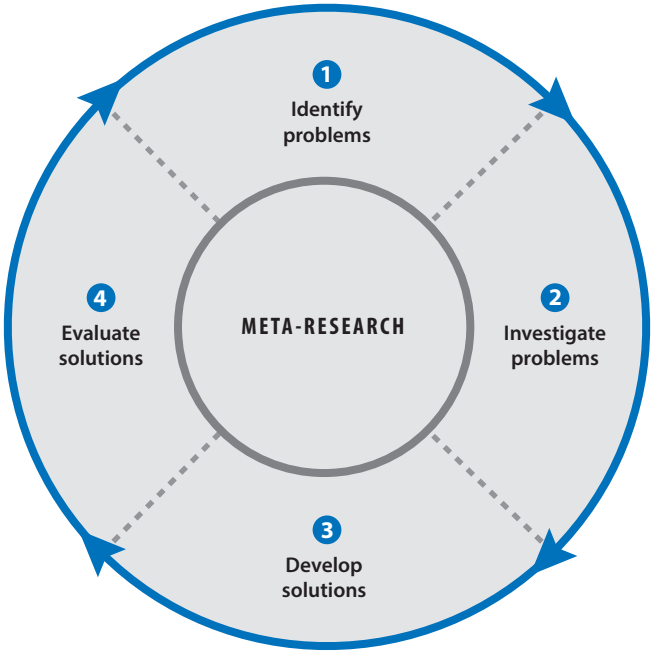

Figure 1

A translational framework for meta-research. Adapted with permission from Tom Hardwicke and available at **https://osf.io/cqrp8/under** a CC-BY4.0 license.

1.3. A Translational Framework for Meta-Research

In this review, we map the endeavor of meta-research using the translational framework depicted in **Figure 1**. This framework is not necessarily comprehensive and aims to be descriptive rather than prescriptive. The goal is to highlight how individual meta-research projects form part of a broader effort to continuously calibrate the scientific ecosystem toward higher standards of efficiency, credibility, and quality. In the first stage, researchers identify potential problems in the scientific ecosystem, such as publication bias or suboptimal research design. This stage is based on conceptual development and theoretical argumentation, although occasionally modeling, simulations, or case demonstrations are employed. In the second stage, researchers conduct empirical studies to investigate the prevalence and severity of the proposed problems, often by scrutinizing the published literature or surveying working scientists. In the third stage, potential solutions are generated and implemented, such as the development of new infrastructure, policy changes by key stakeholders (e.g., universities, journals, or funders), or development of educational programs. In the final stage, researchers evaluate the effectiveness of potential solutions, either under controlled conditions or in the wild as they are deployed. Ideally, later stages may also reciprocally inform earlier stages, creating feedback loops. In this review, we visit each of the stages of the framework, providing illustrative examples and discussing methodological challenges.

2. IDENTIFYING PROBLEMS

Researchers may identify issues that could potentially cause inefficiency, hamper research quality, and undermine the veracity of the published literature. This can be based on theoretical arguments, modeling, simulations, or early empirical data (e.g., case demonstrations or limited

evaluation of studies suffering from these problems). A central challenge is the sheer complexity of the scientific ecosystem. Multiple stakeholders, infrastructures, and processes push and pull in different directions, interact, and evolve. Any individual problem, even if it can be reasonably well described, forms part of a complex causal network of interleaved factors. Delineating symptoms versus causes is typically not straightforward. For example, one assessment mapped 235 different biases in the biomedical literature (Chavalarias & Ioannidis 2010), and biases may manifest differently or have different prevalence and consequences across distinct scientific fields (Fanelli et al. 2017, Goodman 2019). Other discipline-specific characteristics, such as the ratio of true (non-null) to absent (null) relationships among those relationships under scrutiny, and design considerations, such as statistical power and flexibility in design and analysis decisions, may translate to very different chances of getting correct or wrong answers (Ioannidis 2005). In this section, we discuss some salient problems that have been proposed, debated, and studied.

2.1. Incentives and Norms

Many problems may arise in the scientific ecosystem due to a fundamental misalignment of scientific ideals and incentive structures. While there is no doctrine defining a set of scientific ideals, the sociologist Robert Merton suggested that scientists share a set of informal cultural norms (Merton 1973): universalism (researchers should evaluate claims based on the evidence rather than irrelevant personal characteristics such as ethnicity, nationality, gender, or professional affiliation); communalism (scientific methods and results belong to the entire scientific community); disinterestedness (science should be free from personal, monetary, and other biases); and organized skepticism (researchers should engage in impersonal critical scrutiny). The extent to which these norms accurately describe scientists' behaviors and beliefs has been challenged (Mulkay 1976), although some survey evidence suggests that many working scientists subscribe to them (Anderson et al. 2010).

Several authors have asserted that key stakeholders in the scientific ecosystem are acting counter to these norms by exerting a preference for scientific findings with certain aesthetic qualities at the expense of authenticity (Giner-Sorolla 2012, Nosek et al. 2012). Specifically, stakeholders such as universities, journals, reviewers, and funders may prefer newsworthy, positive, or clean findings over incremental, negative, or messy findings. Because they exert considerable influence over how research is performed and evaluated, these stakeholders could be creating selection pressures that affect the quality and veracity of research.

Several modeling and simulation efforts have explored the consequences of incentive structures by mimicking the workings of the reward system in science (Bakker et al. 2012, Grimes et al. 2018, Higginson & Munafò 2016, Smaldino & McElreath 2016). For example, Grimes et al. (2018) showed how the emphasis on positive results in top-tier journals can undermine the trustworthiness of scientific findings. Conversely, trustworthiness improved when journals were agnostic as to whether a result was positive or not. The authors' model also suggested that a decrease of allocated funding amplifies competition between scientists, giving rise to an environment in which false-positive results are actively rewarded, thereby further decreasing trustworthiness. Smaldino & McElreath (2016) designed a dynamic model of competing research labs to demonstrate the "natural selection of bad science." Successful labs pursue novel, positive results by selecting suboptimal methods that can deliver such results in large quantities. Ultimately these labs receive higher payoffs (e.g., citations, prestige, funding), allowing them to reproduce at a higher rate and replicate through many offspring labs, perpetuating suboptimal methods in the scientific ecosystem. Higginson & Munafò (2016) extended this framework to research designs and estimated that a publish-or-perish culture contributes to designs with inadequate statistical power and high false-positive rates.

2.2. Lack of Transparency

The importance of transparency has a long historical precedent—the world's oldest scientific institution, the Royal Society, has the motto *nullius in verba*, "take nobody's word for it." Transparency is also regarded by some as an ethical imperative. For example, the Declaration of Helsinki on ethical principles for clinical research states that researchers have an ethical obligation to publish and disseminate complete and accurate reports of their research (World Med. Assoc. 2013). A lack of transparency in the conduct, reporting, and dissemination of research may undermine trust in science (Vazire 2017), waste resources (Chalmers & Glasziou 2009), and disrupt self-correction mechanisms (Ioannidis 2012).

One major concern—publication bias—refers to the phenomenon whereby scientific findings are selectively published based on various aesthetic characteristics of the results, such as being positive or newsworthy (Dwan et al. 2013). Publication bias can emerge through multiple mechanisms. For example, it could be that journals (or reviewers) express a preference for certain types of scientific findings and selectively publish findings that best match those preferences. Alternatively or additionally, researchers may not submit findings with certain characteristics for publication (the file drawer effect; Rosenthal 1979). Publication bias can also emerge from multiple sources, including selective reporting of entire studies or experiments, individual experimental conditions, specific measured outcomes, and outcomes arising from particular analyses (Phillips 2004). The consequence is that publications in academic journals fail to capture all of the findings generated by the scientific enterprise, thus providing a skewed impression of the evidentiary landscape.

Even when research findings are reported, they can be undermined by a lack of transparency about how they were generated. Much activity in the scientific ecosystem centers on research articles as a principal commodity (Young et al. 2008); however, a research article is an incomplete snapshot that cannot fully capture the rich network of research resources (e.g., protocols, materials, raw data, analysis scripts) that provide a more direct account of the research process and output (O. Klein et al. 2018). Being able to access research resources can facilitate independent verification and evidence synthesis, and it may promote new discoveries. For example, access to raw data could enable reanalyses that probe the reproducibility and robustness of the original findings (Hardwicke et al. 2018), facilitate more sophisticated forms of meta-analysis (Tierney et al. 2015), or generate new insights through the application of novel techniques or merging with other data sets (Voytek 2016). Occasionally, overriding ethical or legal concerns may limit transparency (Meyer 2018). In such cases, explicitly declaring such negative constraints on sharing should be a minimum expectation (Morey et al. 2016).

Without transparency it can be unclear what the original research hypothesis was, what the raw data looked like, how many studies or analyses were attempted, and how many unattractive results were disregarded. This information is indispensable for properly appraising research.

2.3. Statistical Schools of Thought and Statistical Misuse

Most scientific claims are reinforced by a scaffold of statistical analyses that support inductive inferences from samples of data. There are multiple approaches to statistical inference, including Bayesian and likelihood-based, but frequentist inference is the most prevalent (Chavalarias et al. 2016). Deep philosophical rifts exist between these schools of thought (Mayo 2018). Frequentist inference is often used in the form of null-hypothesis significance testing (NHST), a hybrid of two different statistical schools of thought (Gigerenzer 2004, Goodman 1993) that is highly prone to misinterpretation and misuse (Szucs & Ioannidis 2017b, Wagenmakers 2007).

Regardless of the statistical paradigm employed, all statistical analyses have the potential to be misused. There are many researcher's degrees of freedom in data analysis and interpretation—just

10 binary analytic choices result in $2^{10} = 1{,}024$ unique analysis specifications (Gelman & Loken 2014). This flexibility can lead to large vibration of effects whereby many different results can be obtained from the same data and research question; results from the same study may often point in opposite directions (the so-called Janus phenomenon) (Ioannidis 2008, Patel et al. 2015). This situation can easily be exploited, whether intentionally or not, to extract more desirable results (see Section 2.1) from any given data set.

Making analytic decisions in a data-dependent manner without using appropriate corrections generates more false positives (Simmons et al. 2011). A series of data-dependent activities, collectively known as *p*-hacking, include stopping data collection, dropping outliers, selecting covariates, or inappropriately rounding *p*-values based on whether those actions shift the results toward statistical significance.[1] Analytic flexibility varies between domains and partly depends on the degree of standardization of data processing and analytic approaches. In the field of neuroimaging, for example, there is enormous flexibility in the data processing pipeline (Carp 2012), a situation that enabled one research team to detect brain activity in a dead Atlantic salmon (Bennett et al. 2009).

A related phenomenon that can complicate *p*-hacking is opaque HARKing (hypothesizing after results are known): presenting a finding as if it were hypothesized all along, thus adding false confidence in its validity (Kerr 1998, O'Boyle et al. 2013). HARKing adds further degrees of freedom to the analysis process, enabling *p*-hacked findings to be explained convincingly by tailor-made hypotheses.

2.4. Reproducibility

Use of the term "reproducibility" and related terms such as "replicability" and "repeatability" can vary across fields (Barba 2018). Goodman et al. (2016) proposed one framework that delineates methods reproducibility (obtaining similar results given the same data and analytical tools; often called analytic reproducibility or computational reproducibility), results reproducibility (obtaining similar results given the same analytical and experimental tools but new data; often called replication), and inferential reproducibility (drawing qualitatively similar inferences from an independent methods or results reproduction of a study).

Reproducibility is a core tenet of the scientific method: If one researcher performs a study and makes a claim, a second researcher should be able to repeat the original methods and obtain similar results. Repeating original analyses with the raw data should enable recovering the originally reported findings (Hardwicke et al. 2018). However, in the context of stochastic phenomena, we should expect the findings of replication studies to differ to some extent from original studies (Stanley & Spence 2014). It is also unclear how (or if) one should compare the results of two studies (an original and a replication) and conclude whether one was successful at replicating the other (Goodman et al. 2016, Nosek & Errington 2017). Some replication attempts have sparked heated debates that often result in further nonreproducibility of inferences: Despite examining the same results, researchers can disagree about what they mean (Ioannidis 2017).

3. INVESTIGATING PROBLEMS

At this stage, researchers conduct more in-depth empirical investigations to examine the prevalence and severity of problems. Investigations may involve meta-epidemiological assessments of potential bias, the impact of study characteristics on observed effects, the distribution of research

[1] To hack your own way to scientific glory, see **https://projects.fivethirtyeight.com/p-hacking/**.

evidence in different settings, or quantification of heterogeneity (Murad & Wang 2017). Most studies adopt retrospective observational designs and typically involve manual examination of published research articles. A serious challenge is that low transparency may undermine efforts to systematically study other problems. Consequently, some meta-epidemiological studies rely on surrogates that assess whether the pattern of published results is compatible with the theoretical impact of some particular bias. Because a number of benign factors may also contribute to such patterns, such surrogates are approximate/imperfect indicators of bias.

3.1. Incentive Structures

In evaluations for hiring, promotion, and tenure, many institutions use simple metrics, such as the journal impact factor, which are known to be problematic (Moher et al. 2018). Survey evidence suggests that many scientists view number of publications, journal ranking, and authorship order as being directly associated with performance assessment and career promotions (van Dalen & Henkens 2012, Walker et al. 2010). In some countries, including China, South Korea, and Turkey, cash rewards are offered for publishing in top-tier journals (Franzoni et al. 2011). Another strong incentive is to claim novelty, which many biomedical researchers do even when this is demonstrably false. In a study of 1,101 randomized controlled trials (RCTs) with 5 or more preceding trials combined in a meta-analyses, 46% cited only 0 or 1 prior trial on the subject, a percentage that increased when there were more prior trials to cite (Robinson & Goodman 2011). Such incentives may hinder research quality as competitive environments with greater publication pressure are more likely to report statistically significant results, a pattern that could be indicative of biased reporting (Fanelli 2010). The need to align incentive structures with good scientific practice is now widely recognized. For example, in one study, about 80% of surveyed scientists thought that incentivizing better research practices would improve reproducibility (Baker 2016).

3.2. Publication Bias and Selective Reporting

Empirical investigation of publication bias is challenging because unpublished studies and results are difficult to unearth. One approach is to seek out signals of publication bias in the published literature. For example, Fanelli (2011) manually examined a sample of 4,656 articles across scientific disciplines and observed an overwhelming frequency of positive (i.e., statistically significant) findings (see also Sterling 1959). Similarly, text-mining extraction of p-values from MEDLINE abstracts and full-text articles in PubMed Central showed that 96% of abstracts and full-text articles claimed significant results with p-values <0.05 (Chavalarias et al. 2016). This is simply too good to be true; it is implausible that scientists are routinely testing hypotheses that so frequently turn out to be accurate, especially when statistical power is typically very low (Ioannidis & Trikalinos 2007).

More direct evidence for publication bias has arisen from comparing public records, such as dissertations or study registries, with the published literature. For example, in the domain of management research, O'Boyle and colleagues (2013) found that the ratio of supported to unsupported hypotheses was more than twice as high in a corpus of published articles when compared to corresponding student dissertations. Similarly, Franco and colleagues (2014) capitalized on an institutional rule that required questionnaires and data underlying a series of psychology studies to be made publicly available. By comparing these materials to published research articles, they observed that about 40% of articles did not report all experimental conditions, about 70% of articles did not report all outcome variables, and reported effect sizes were about twice as large as unreported effect sizes (also see Franco et al. 2016).

In medicine, researchers have used study protocols available in ethics board documentation or study registries to identify selective reporting of studies and outcomes in the published literature (Chan et al. 2004, Dechartres et al. 2017, Dickersin et al. 1992, Dwan et al. 2011, Easterbrook et al. 1991, Goldacre et al. 2019, Ross et al. 2012). For example, a systematic review of studies comparing those sources of information with their corresponding trial reports highlighted recurrent discrepancies (Dwan et al. 2014).

3.3. Transparency of Research Resources

Assessment of articles published across multiple scientific domains suggests minimal availability of critical research resources such as raw data, protocols, materials, and analysis scripts (Alsheikh-Ali et al. 2011, Hardwicke et al. 2019b). Wallach et al. (2018) assessed a random sample of 149 articles published in the biomedical domain between 2015 and 2017 and found that 19 articles had data availability statements, 31 articles had materials availability statements, one article shared a full protocol, and no articles shared analysis scripts. Data sharing may have improved recently in some domains, but there is still much room for improvement (compare Iqbal et al. 2016, Wallach et al. 2018).

Attempts to request research resources directly from researchers, particularly raw data, are often unsuccessful (Naudet et al. 2018, Rowhani-Farid & Barnett 2016, Vanpaemel et al. 2015, Vines et al. 2014, Wicherts et al. 2006), even for some of the most influential studies (Hardwicke & Ioannidis 2018b). For example, Hardwicke & Ioannidis contacted the authors of 111 highly cited studies published in psychology and psychiatry between 2006 and 2016 and asked if they would be willing to share the associated raw data. Only 15 data sets (14%) were made available in a completely unrestricted form, and ultimately data from 76 studies (68%) were not made available in any form. Naudet et al. (2018) were more successful and obtained 17 data sets from a sample of 37 (46%) RCTs published in the *BMJ* and *PLOS Medicine* for the purpose of reanalysis.

More extensive assessment, continuous monitoring, and evaluation of research resource transparency are limited by the time-consuming nature of this type of research, which typically requires manual data extraction and coding. It is possible that computational tools can be developed to automatically extract similar information; however, the performance of such tools will need careful assessment to ensure reasonable sensitivity and specificity.

3.4. Suboptimal Research Design

Suboptimal research design may produce misleading results, wasting already scarce resources (Ioannidis et al. 2014). Numerous biases in research have been shown to affect the scientific literature (Fanelli et al. 2017), making it difficult to detect small effect sizes. Deficiencies differ from one scientific field to another, but some patterns are highly prevalent across fields. For example, lack of sufficient power to detect a range of plausible effect sizes is common across many different disciplines (Button et al. 2013, Cohen 1962, Ioannidis et al. 2017b, Moher et al. 1994, Sedlmeier & Gigerenzer 1989, Smaldino & McElreath 2016, Szucs & Ioannidis 2017a). Furthermore, small studies tend to generate more heterogeneous results (IntHout et al. 2015). For example, a meta-epidemiological study assessing 85,002 forest plots from the Cochrane Database of Systematic Reviews showed that most large treatment effects originated from small studies, and when aggregated with other studies, pooled treatment effects tended to be smaller (Pereira et al. 2012).

Other design considerations have also been associated with effect magnitude. For example, exaggerated treatment estimates are observed (on average) in observational studies compared to RCTs (Hemkens et al. 2016), surrogate outcomes compared to patient relevant outcomes (Ciani

et al. 2013), single-center compared to multicenter clinical trials (Dechartres et al. 2011), and studies with inadequate allocation concealment, random-sequence generation, and blinding (Page et al. 2016).

3.5. Statistical Misuse

Intentional or unintentional misuse of statistics may occur during the selection, implementation, reporting, and interpretation of statistical analyses (**Table 1**). For example, a recent meta-epidemiological survey found that most RCTs with subgroup claims did not adjust for multiple testing, did not use an appropriate test of interaction, and were rarely validated (Wallach et al. 2017). Similarly, a study of 157 neuroscience articles found that 79 made interaction claims but did not appropriately test for one (Nieuwenhuis et al. 2011); instead, they inferred interaction when the outcome in one group was statistically significant and the outcome in the other was not, a known statistical fallacy (Gelman & Stern 2006). The causes of statistical misuse are multifaceted. The misapplication of statistical tools may reflect a response to a flawed incentive structure (Section 2.1) or may be due to a lack of understanding and/or poor training, leading to repeated use of mindless statistical rituals (Gigerenzer 2004).

3.6. Reproducibility

Replication studies remain uncommon in many fields (Hardwicke et al. 2019b, Iqbal et al. 2016, Makel et al. 2012, Sterling 1959, Wallach et al. 2018). In recent years however, a spate of high-profile replication attempts in psychology (Open Sci. Collab. 2015, Yong 2012) and industry-based preclinical research (Begley & Ellis 2012, Prinz et al. 2011) have generated serious concern about a transdisciplinary reproducibility crisis (Baker 2016). A series of multilaboratory efforts adopting high transparency standards and (typically) relatively large sample sizes have been deployed to investigate replicability across several fields. The Reproducibility Project: Psychology (RPP), for example, set out to replicate 100 studies published in high-profile journals. The project reported several indices of replication success; for example, although 97 of the original studies had statistically significant results ($p < 0.05$), only 36 of the replications did so. When the outcome of a replication study appears to contradict previous evidence, it is important to consider several factors, such as the track record of the theory under scrutiny and the fidelity of the replication attempt. Some research has explored whether prediction markets can be used to estimate replication success (see the sidebar titled Estimating Replication Success Using Prediction Markets and Surveys).

One fairly consistent pattern that has emerged from the RPP and subsequent large-scale replication studies in the social sciences is that effect sizes observed in the replication studies have been on average approximately half as large as those reported in the original studies (Camerer et al. 2016, 2018), consistent with the idea that most published effects are inflated by selective reporting and other biases (Ioannidis 2005, 2008). A Bayesian analysis of the RPP project also highlighted that the evidential value of many of the original studies (and some of the replication studies) was too weak to support robust inferences (Etz & Vandekerckhove 2016).

Several studies have also investigated whether published findings can be recovered by repeating the original analysis on the raw data (i.e., analytic reproducibility; e.g., Hardwicke et al. 2018, Naudet et al. 2018, Stodden et al. 2018). For example, Hardwicke et al. (2018) encountered at least one nonreproducible value in 24 out of 35 published psychology articles, often due to ambiguous, incomplete, or incorrect specification of the original analyses or mismanagement of data files. Some issues were resolved after the original authors provided (previously unreported)

Table 1 Illustrative cases of statistical misuse affecting the selection, implementation, reporting, and interpretation of analyses reported in the published literature

Selection issues	Illustrative references
Using inappropriate statistical models (e.g., using metric models to analyze ordinal data, using an independent *t*-test in a repeated-measures design, neglecting model assumptions)	Ernst & Albers 2017, Liddell & Kruschke 2018, Strasak et al. 2007
Circular analysis (e.g., attempting to correlate brain activity measures with personality measures after selecting from the former only data that have surpassed a threshold; also known as double dipping)	Fiedler 2011, Vul et al. 2009
Implementation issues	**Illustrative references**
Failure to account for multiplicity (e.g., multiple comparisons, optional stopping)	Armitage et al. 1969, Cramer et al. 2016, John et al. 2012, Strasak et al. 2007, Wallach et al. 2017
Exploiting flexibility in analysis decisions in order to obtain more favorable outcomes	Carp 2012, Orben & Przybylski 2019, Silberzahn et al. 2018, Steegen et al. 2016
Unjustified outlier exclusion (e.g., excluding data points ad hoc in a way that makes outcomes more favorable to the hypothesis under scrutiny)	Bakker & Wicherts 2014, John et al. 2012
Falsification of data	Carlisle 2012, Fanelli 2009, John et al. 2012, Simonsohn 2013
Reporting issues	**Illustrative references**
Inconsistencies (e.g., the reported combination of degrees of freedom, test statistic, and *p*-value are incompatible, or the reported combination of means for integer data, sample size, and number of items is incompatible)	Bakker & Wicherts 2011, Brown & Heathers 2017, Nuijten et al. 2016
Misleading or suboptimal graphical presentation (e.g., inappropriate truncation of the *y*-axis, misidentification or nonidentification or error bars, not representing distributional information)	Lane & Sándor 2009, Weissgerber et al. 2015
Incomplete or unclear design or analysis specification (e.g., not identifying statistical procedures, ambiguous description of experimental units, not reporting data exclusions)	Avey et al. 2016, Carp 2012, Glasziou et al. 2014, Hardwicke et al. 2018, Lazic et al. 2018, Strasak et al. 2007, Vasilevsky et al. 2013, Weissgerber et al. 2018
Incomplete reporting of outcomes (e.g., not reporting effect sizes, interval estimates, or standard deviations)	Counsell & Harlow 2017, Cumming et al. 2007
Distorted presentation of nonsignificant results (e.g., spins in conclusions)	Boutron et al. 2010
Selective reporting (e.g., only reporting experiments, outcomes, or analyses that achieved statistical significance)	Chan et al. 2004; Dwan et al. 2011, 2013, 2014; Franco et al. 2014, 2016; Goldacre et al. 2019; John et al. 2012; Easterbrook et al. 1991
Presenting post hoc hypotheses as if they were specified a priori (opaque HARKing)	John et al. 2012, Kerr 1998, Wagenmakers et al. 2012
Interpretation issues	**Illustrative references**
Incorrectly assuming that a nonsignificant outcome means that there is no effect	Fidler et al. 2006, Hoekstra et al. 2006, Schatz et al. 2005, Sedlmeier & Gigerenzer 1989
Assuming that the difference between significant and not significant is itself significant, or relatedly, erroneous analysis of interactions	Gelman & Stern 2006, Nieuwenhuis et al. 2011, Wallach et al. 2017

Based on table 1 in Hardwicke et al. (2019a). Abbreviation: HARKing, hypothesizing after results are known.

ESTIMATING REPLICATION SUCCESS USING PREDICTION MARKETS AND SURVEYS

Prediction markets elicit group beliefs about replication success by having participants place monetary bets (Camerer et al. 2016, 2018; Dreber et al. 2015; Forsell et al. 2018). Dreber et al. (2015) found that prediction markets (71%) outperformed premarket surveys (58%) when participants were asked to predict replication success for 44 studies from the RPP. However, subsequent studies evaluating social science studies observed that market beliefs were not more accurate than surveys (Camerer et al. 2016, 2018). A recent study (Forsell et al. 2018) of the Many Labs 2 replication project (R.A. Klein et al. 2018) reported that performance depended on the type of replication outcomes being predicted. Prediction markets more accurately predicted replication significance (i.e., a statistically significant effect in the same direction as the original study), whereas surveys more accurately predicted replication effect sizes. The relatively low cost and rapid results of prediction markets and surveys make them attractive tools for predicting replication outcomes. However, in existing assessments, forecasters had prior information about the studies under scrutiny. Thus, a future challenge for market predictions will be their performance with new, unfamiliar studies.

information that enabled reproducibility. Importantly, there was no clear evidence that the conclusions of the original studies had been undermined. Generally, studies of analytic reproducibility have highlighted that basic human error is common, suggesting that greater attention should be paid to quality control systems and use of software tools that enable writing reproducible scientific papers (O. Klein et al. 2018, Marwick et al. 2017).

4. DEVELOPING SOLUTIONS

During this stage researchers and other stakeholders such as universities, funders, and journals, attempt to develop and implement solutions to problems delineated in previous stages in order to improve the efficiency, quality, and credibility of scientific research (Ioannidis 2014, Munafò et al. 2017). Arguably, many reform initiatives are facilitated by the availability of suitable software and technological infrastructure, such as data analysis tools that emphasize reproducibility and repositories for registering study protocols and sharing critical research resources such as raw data (Spellman 2015). Proper (re)training of the scientific workforce may also be crucial for the success of many such initiatives.

4.1. Journal, Funder, Society, and University Policies

Some journals, funders, academic societies, universities, and other institutions have begun to introduce policy changes trying to address problems. For example, Nosek et al. (2015) developed the Transparency and Openness Promotion (TOP) guidelines, a set of tiered policy recommendations encompassing data sharing, materials sharing, analysis code transparency, data citation standards, design and analysis reporting, preregistration, and replication. At the time of writing, the TOP website (**https://cos.io/our-services/top-guidelines/**) reports that over 5,000 organizations (including journals, publishers, and funders) are signatories; however, this only involves "expressing their support of the principles of openness, transparency, and reproducibility, expressing interest in the guidelines," and a commitment to "conducting a review within a year of the standards and levels of adoption." The website also notes that "We know of over 1,100 journals or organizations that have implemented one or more TOP-compliant policy as of June 2019."

4.2. Preregistration and Registered Reports

Preregistration involves creating a time-stamped, read-only copy of a study protocol (e.g., hypotheses, methods, and analysis plan) and archiving it in a registry before study commencement. The intention is to mitigate or enable detection of questionable research practices, such as *p*-hacking and HARKing, by making clear what was planned and what was not (Nosek et al. 2018). For example, selective reporting can potentially be identified by comparison of the protocol and the report (Goldacre et al. 2019). Additionally, preregistration may help researchers avoid capitalizing on chance by exploiting (perhaps unintentionally) degrees of freedom in the analysis process. Crucially, the intention of preregistration is not to reduce opportunities for exploratory (data-dependent) analyses, but to make clear the exploratory nature of such work (Kimmelman et al. 2014, Wagenmakers et al. 2012). Although the concept of preregistration is emerging in the basic sciences and preclinical domains (Nosek et al. 2018), it has a longer precedent in the context of clinical trials registration (Dickersin & Rennie 2012) and has sparked opposition and debate in some domains (e.g., compare Dal-Ré et al. 2014, Lash & Vandenbroucke 2012).

Implementation of preregistration in practice depends on the research domain, the existence of legal mandates, and the registry that a researcher chooses (or is required) to use (**Table 2**). Some registries are tailored to the needs of specific fields and offer or require completion of specific registration templates. Templates facilitate standardization but may not be optimized for some designs. Registries can also influence the level of transparency conferred by preregistration. Some registries automatically make registrations public (e.g., ClinicalTrials.gov), and

Table 2 Key characteristics of example registries

Registry	Field	Study type	Mandates or incentives	Template(s)	Embargo option	Results reporting
ClinicalTrials.gov	Biomedicine	Clinical trials	FDA regulation (legal) ICMJE statement (for publication)	Yes	No	Mandatory for some trials
American Economic Association RCT Registry	Economics	RCTs	American Economic Association journals (for publication)	Yes	No	No
AsPredicted	Social sciences	General	None declared	Yes	Yes	No
Open Science Framework	All	All	Preregistration Challenge (cash incentives)	Optional	Yes	Optional
PROSPERO	Biomedicine	Systematic reviews	UK National Institute of Health Research (for funding)	Yes	No	Optional
WHO International Clinical Trials Registry Platform[a]	Biomedicine	Clinical trials	Country specific regulation (legal) ICMJE statement (for publication)	Mixed	Mixed	Mixed

[a]WHO database containing multiple registries, each with their own specific features.
Abbreviations: FDA, US Food and Drug Administration; ICMJE, International Committee of Medical Journal Editors; RCT, randomized controlled trial; WHO, World Health Organization.

some offer an optional and time-limited embargo period during which the registration is hidden before the information eventually becomes public (e.g., the Open Science Framework). Others allow registrations to be kept hidden indefinitely (e.g., AsPredicted), which may help to allay concerns about ideas being scooped. However, hidden registrations cannot be effectively monitored by the scientific community. Monitoring can help address issues such as creating multiple similar registrations, registering but not publishing, and registering but still using questionable research practices (Ioannidis et al. 2017a).

The registered reports article format involves embedding protocol registration directly within the publication pipeline (Chambers 2013). Study protocols are peer reviewed and may be offered in-principle acceptance for publication before the study has even been conducted. By focusing on the quality of study design rather than the aesthetic appeal of the study findings, registered reports may improve study quality and mitigate publication bias. This type of publishing model appears to have been employed by the *European Journal of Parapsychology* from 1976 for almost two decades (Wiseman et al. 2019) as well as *The Lancet* from 1997 for at least a decade (Hardwicke & Ioannidis 2018a). The current registered reports format (**http://cos.io/rr**) was introduced at *Cortex* in 2013 (Chambers 2013), and adoption has spread across journals and disciplines (Hardwicke & Ioannidis 2018a).

4.3. Reporting Guidelines

Reporting guidelines are intended to draw attention to key design and analysis decisions and ensure they are adequately reported in research reports (Altman & Simera 2016). In 2008 the EQUATOR (Enhancing the Quality and Transparency of Health Research) network was launched to provide resources, education, and training to facilitate good research reporting (**https://www.equator-network.org/**). Currently the network provides 411 reporting guidelines covering a variety of study types (**Table 3**), but it is predominantly focused on health research.

4.4. Peer Review

Peer review is often considered to be an essential quality control gateway that should prevent low-quality research from entering the scientific literature. There are many studies on peer review, many of which are presented at the International Conference on Peer Review that runs every four years. However, assessment of interventions to improve peer review has been relatively sparse (Bruce et al. 2016). Several variations or amendments to peer review procedures have been proposed to make it more effective and mitigate potential negative consequences, such as the introduction of bias by peer reviewers themselves. For example, there has been much debate and some empirical scrutiny of whether the identities of peer reviewers and/or the content of their reviews should be made publicly available (Ross-Hellauer & Görögh 2019) and whether peer review procedures should be double-blind (Justice et al. 1998, McGillivray & De Ranieri 2018). The registered reports publication model represents a radical departure from traditional peer review procedures, as it involves results-blind peer review. It has also been proposed that peer reviewers might decline to review a particular manuscript if critical research resources such as raw data, materials, and analysis scripts are not made publicly available or the manuscript does not contain a statement explaining why they cannot be made available (Morey et al. 2016). Finally, there has been discussion about whether specialized statistical review (a process that appears to have had some degree of success in biomedical domains) might also help other fields where statistical review appears to be relatively rare, such as psychology (Hardwicke et al. 2019a).

Table 3 Some widely used reporting guidelines available on the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network

Study type	Guidelines	Current version	*N* citations[a]
Randomized trials	CONSORT: CONsolidated Standards Of Reporting Trials	2010[b]	6,439
Observational studies	STROBE: Strengthening the Reporting of Observational Studies in Epidemiology	2008	8,653
Systematic reviews	PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses	2009	34,295
Study protocols	SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials	2013	997
	PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols	2015	2,524
Diagnostic/prognostic studies	STARD: STAndards for Reporting Diagnostic accuracy studies	2015[b]	652
	TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis	2015	821
Case reports	CARE: CAse REport guidelines	2013	369
Clinical practice guidelines	AGREE: Appraisal of Guidelines, Research and Evaluation	2016	65
	RIGHT: Reporting Items for practice Guidelines in HealThcare	2017	35
Qualitative research	SRQR: Standards for Reporting Qualitative Research	2014	420
	COREQ: COnsolidated criteria for REporting Qualitative research	2007	3,352
Animal preclinical studies	ARRIVE: Animals in Research: Reporting In Vivo Experiments	2010	2,120
Quality improvement studies	SQUIRE: Standards for QUality Improvement Reporting Excellence	2015[b]	223
Economic evaluation	CHEERS: Consolidated Health Economic Evaluation Reporting Standards	2013	732

[a]Scopus citation numbers as of April 10, 2019, for the guideline publications (including publications appearing simultaneously in several journals). If relevant, citation counts are shown for the latest version only.
[b]Date refers to publication of an updated version of the guideline.

4.5. Collaboration

Pooling expertise, financial resources, and other resources in collaborative work may improve statistical power, reproducibility, access to unique populations, and results generalizability (Ioannidis 2014, Munafò et al. 2017). Collaborative efforts have completely transformed many scientific fields, such as genetics (Seminara et al. 2007), and have become the norm in many physical and space sciences. This model has recently started to gain traction in disciplines where such large-scale collaboration was previously rare, such as psychology (Open Sci. Collab. 2015, R.A. Klein et al. 2018). The new Psychological Science Accelerator operates as a large network of labs with different committees responsible for coordinating various tasks, such as study selection and data management (Moshontz et al. 2018). Similarly, the Observational Health Data Sciences and

Table 4 Various proposed solutions for improving statistical inference on a large scale

Proposed solution	Application to past publications	Application to future publications
Lower *p*-value thresholds	A rather simple temporizing solution	Has potential collateral harms and success depends on adoption or enforcement by stakeholders (e.g., journals, funders, societies)
Abandon *p*-value thresholds and instead use exact *p*-values	Many published *p*-values have only been reported with thresholds.	Success depends on extent of adoption or enforcement by stakeholders
Abandon *p*-values entirely	Not easy because insufficient information might be available to compute other statistics; many articles do not report effect sizes and/or confidence intervals	Previous efforts have not gained traction. May be more successful in some fields (e.g., assessment of diagnostic performance or choosing predictors for prognostic models in which *p*-values would make little sense)
Focus on effect sizes and their uncertainty	Often this information is not reported at all, but it has become more common in recent literature.	Relevant to the vast majority of the clinical literature; should be heavily endorsed as more directly linked to decision making and may be easier to promote than more sophisticated solutions
(Re)train the scientific workforce	Takes time and major commitment to achieve sufficient statistical literacy	Potentially a more effective solution in the long-term but may require major recasting of training priorities in curricula
Address biases that lead to inflated results	Requires major training; biases are often impossible to detect in published reports	Preemptively dealing with biases is ideal but needs concerted commitment of multiple stakeholders to promote and incentivize better research practices

Adapted from Ioannidis (2018b).

Informatics initiative aims to bring collaborators from multiple sites together in order to merge health data from multiple sites into one standardized database, reproduce analyses across sites, and explore the optimization of design and analytic choices in observational research (Madigan et al. 2014).

Overall, an increase in large-scale collaborations seems welcome; however, it does raise practical and conceptual issues. For example, accurately and effectively crediting hundreds of authors in a team effort could prove difficult in the current system, which values certain authorship positions more than others. Large-scale collaborations may also dilute beneficial competition and disagreement by focusing on finding a lowest common denominator consensus at the expense of more radical approaches. One interesting hybrid approach known as adversarial collaboration involves two groups of researchers who have a theoretical disagreement collaborating on a project in an effort to maximize the informational value of study design and minimize the influence of their respective biases (Matzke et al. 2015). Some empirical work is already addressing the relative merits of small versus large team science (Wu et al. 2019).

4.6. Statistical Reform

In response to widespread statistical misuse (see Section 3.5), there have been many proposals for statistical reform (see **Table 4**). The *American Statistician* recently published a collection of 43 articles containing several such proposals (Wasserstein et al. 2019). As NHST is the dominant statistical paradigm in most disciplines (Section 2.3), many solutions are focused on either trying to improve the use of NHST or completely replace it. For example, Benjamin et al. (2017) proposed that reducing the traditional threshold for declaring statistical significance from 0.05

to 0.005 would help researchers more appropriately calibrate their inferences to the strength of the evidence. It has also been suggested that, instead of using a theoretical null distribution, researchers use a data-driven null constructed using negative controls (i.e., associations that we know should be null) to produce calibrated *p*-values (Schuemie et al. 2014). Others have proposed completely abandoning significance testing (McShane et al. 2019). However, an empirical assessment of articles published during a journal ban of significance testing suggested a tendency to overstate conclusions (Fricker et al. 2019), which is a concern in the absence of any statistical inference (Ioannidis 2019).

Others suggest a switch to alternative paradigms, such as Bayesian statistics, which has several advantages (Wagenmakers 2007). However, it is noteworthy that different inferential approaches often arrive at similar conclusions, at least in simple scenarios (van Dongen et al. 2019). Furthermore, the success of any given approach is largely dependent on the capabilities of the researchers conducting the statistical analyses. Statistical reform is therefore heavily dependent on resolving other issues, such as poor statistical education (Altman 1994, Goodman 2019), misaligned incentives (Nosek et al. 2012), and a lack of transparency, which complicates independent evaluation. Innovative approaches may be worth evaluating, such as requiring authors to indicate their degree of belief in their findings (Goodman 2018).

4.7. Evidence Synthesis

The Cochrane Collaboration was founded in 1993 and quickly became a champion of evidence synthesis. Several more recent initiatives have been launched in order to encourage a more transparent and collaborative process for evidence synthesis such as the Stanford MetaLab (**http://metalab.stanford.edu/**) in developmental psychology and PROSPERO (**https://www.crd.york.ac.uk/prospero/**) for preregistration of systematic reviews. Systematic reviews and meta-analyses are growing exponentially, with an increase of 2,728% and 2,635%, respectively, between 1991 and 2014, but many are redundant, misleading, or conflicted (Ioannidis 2016).

5. EVALUATING SOLUTIONS

The effectiveness of reform initiatives will depend not only on their theoretical sophistication, but on how well they are implemented. Evaluation and ongoing monitoring of reform initiatives are crucial to detect unintended negative consequences and verify that anticipated benefits are being realized in practice (Ioannidis 2015). The success of an intervention depends on how well it is calibrated to the needs, motivations, and capability of scientists, and these factors may vary substantially across research communities. Furthermore, any single reform initiative will typically only address a subset of a complex range of overlapping and interacting processes that are influenced by multiple stakeholders within the scientific ecosystem. Even with best intentions and flawless implementation, reform initiatives may fail because the system evolves to resist them.

Conducting informative evaluation studies can be challenging. Many initiatives are introduced without considering evaluation, which means that most evaluation studies are necessarily limited to retrospective observational designs. A given reform initiative may involve both proximal and distal goals that will emerge across the short, medium, and long terms. Distal goals may take longer to materialize and be more difficult to isolate and operationalize, but they are critical indices of a reform initiative's success or failure. Proximal outcomes can be measured sooner, providing feedback that can be used to address weaknesses and optimize reform initiatives. Here, we describe

some specific examples of evaluating solutions relevant to journal policy, reporting guidelines, preregistration, and registered reports.

5.1. Journal Policy

Generally, policy effectiveness appears to depend on how stringent policy requirements are, how the policy is worded and interpreted, and how robustly the policy is enforced. For example, data sharing policies that recommend authors to share upon request tend to be markedly less effective relative to more stringent policies that require authors to make data publicly available in an online repository prior to publication (Hardwicke et al. 2018, Nuijten et al. 2017, Rowhani-Farid & Barnett 2016). For instance, Nuijten et al. (2017) observed a dramatic increase in data availability from 8.6% to 87.4% of articles after a new policy at *Judgement and Decision Making* asked authors to publicly share data prior to article publication. By contrast, data sharing at a comparator journal with no data sharing policy, the *Journal of Behavioral Decision Making*, remained negligible across the same period (prepolicy, 0%; postpolicy, 1.7%).

Although some data sharing policies might achieve their proximal goal of increasing data availability, they are not necessarily achieving their more distal goal of facilitating data reuse or results verification. Using an interrupted time series analysis, Hardwicke et al. (2018) observed a substantial increase in data available statements after a mandatory data sharing policy was introduced at the journal *Cognition* (from 25% of prepolicy articles to 78% of postpolicy articles). However, among data sets that were reportedly available, the proportion that were actually available, complete, and understandable was only 22% prepolicy and 62% postpolicy (also see Section 3.6). In response to this study, the *Cognition* editorial team outlined policy changes they intend to implement (Tsakiris et al. 2018)—an example of how meta-research can create a feedback loop between the solution evaluation and solution development stages (**Figure 1**).

5.2. Reporting Guidelines

In line with the recommendations of the International Committee of Medical Journal Editors (ICMJE), medical journals are increasingly adopting reporting guidelines. However, an examination of the online instructions to authors of 168 high-impact-factor journals revealed heterogeneous recommendations on the use of Consolidated Standards of Reporting Trials (CONSORT) guidelines (Shamseer et al. 2016). Sixty-three percent endorsed the use of CONSORT, of which 42% made it a prerequisite for submission. An early pre/post study assessing the impact of CONSORT showed improvements in completeness and transparency of published reports (Moher et al. 2001). For example, unclear description of allocation concealment significantly decreased (mean change −22%). Although the overall quality of reporting improved with uptake of guidelines, it remains suboptimal, and deficiencies persist (Turner et al. 2012).

5.3. Preregistration and Registered Reports

In 2005, the ICMJE implemented their policy requiring trial registration for publication, and the number of registrations on ClinicalTrials.gov increased by 73% in 6 months (Zarin et al. 2005). As of April 2019, 301,795 studies have been registered on ClinicalTrials.gov. However, many journals still publish unregistered and retrospectively registered trials (Gopal et al. 2018, Loder et al. 2018, Trinquart et al. 2018). An overview of studies evaluating the registration status of published reports in medical journals found that half of the trials published were not registered and only 20% were prospectively registered (Trinquart et al. 2018).

Registration of trials also intends to prevent selective reporting, yet changes between the registered primary outcome and published outcomes are common (Goldacre et al. 2019, Gopal et al. 2018, Jones et al. 2015, Mathieu et al. 2009, Scott et al. 2015), with up to 31% of articles showing discrepancies in reported versus registered outcomes. Other investigators have reported that mandatory registration of primary outcomes was associated with a substantial decline in the number of trials reporting statistically significant findings, perhaps due to registration effectively mitigating selective reporting (Kaplan & Irvin 2015).

Journal adoption of registered reports has many potential benefits, but a number of implementation issues were found during an exploratory investigation of the format (Hardwicke & Ioannidis 2018a). For example, at the time of the investigation, most registered reports had not been formally registered, and there was no reliable way of tracking their existence and status in the publication pipeline. Furthermore, most in-principle accepted protocols were not publicly available. This investigation is another example of how meta-research can create a virtuous feedback loop to inform and refine solution development (**Figure 1**). Many of the implementation issues identified were quickly addressed (at least in part) through the creation of a central registry for registered reports (**http://cos.io/rr**) and efforts to coordinate and update journal policy to ensure that protocols are registered and made publicly available (Chambers & Mellor 2018). Further evaluation and monitoring will be necessary to ascertain how effective these changes have been and if additional implementation issues should be addressed.

SUMMARY POINTS

1. Meta-research, or research-on-research, is a burgeoning discipline that investigates efficiency, quality, and bias in the scientific ecosystem.
2. We have introduced a translational framework that situates individual examples of meta-research within a broader research agenda that involves (*a*) identifying problems, (*b*) investigating problems, (*c*) developing solutions, and (*d*) evaluating solutions.
3. Using theoretical arguments, modeling, simulations, or early empirical data, meta-researchers have identified potential problems that might cause inefficiency, hamper research quality, and undermine the veracity of the published literature.
4. Empirical investigations have examined the prevalence and severity of problems including publication bias and selective reporting, transparency of critical research resources, suboptimal research design, incentive structures, statistical misuse, and reproducibility.
5. Scientific stakeholders such as universities, funders, journals, and researchers have introduced a number of reform initiatives related to transparency of research resources, preregistration, registered reports, reporting guidelines, peer review, collaboration, statistical reform, and evidence synthesis. The goal is to improve the efficiency, quality, and credibility of scientific research.
6. Evaluation and ongoing monitoring of reform initiatives are crucial to check for unintended negative consequences and verify that anticipated benefits are being realized in practice.
7. Meta-research can help to calibrate the scientific ecosystem toward higher standards by providing a stratum of empirical evidence that informs the iterative generation and refinement of reform initiatives.

FUTURE ISSUES

1. Meta-researchers are operating in the same scientific ecosystem as other researchers and are therefore subject to the same selection pressures that can infuse bias into the research process. It is important that meta-research is held to the same high standards expected of other research.
2. Often researchers promoting reform initiatives are also those with the interest, motivation, and means to evaluate them. These researchers may have the best of intentions, and important studies might not even be performed without their efforts, but such non-independent evaluation does create a risk of bias. Independent evaluations should be prioritized when feasible, and high transparency standards are imperative.
3. As an emerging cross-disciplinary field, meta-research lacks the traditional infrastructures that support more established disciplines. Consequently, many aspects of meta-research are ad hoc and poorly supported, including training, student recruitment, funding, and publication outlets. Career trajectories for meta-researchers are unclear as university departments specializing in meta-research are rare.
4. Meta-research is frequently complicated by a lack of transparency and poor standardization. Important information is often buried in articles and has to undergo time-consuming and error-prone manual extraction. Other information is hidden in journal publishing systems or in researchers' files. Improved transparency standards combined with technological developments will enable more efficient, comprehensive, and effective meta-research.
5. Recent advances in text mining, machine learning, and other automated tools may create new opportunities for meta-research on topics that were previously out of reach. Automated tools may also be able to enhance peer review by detecting simple errors or identifying suggestive patterns that can be evaluated further by a human.
6. Different disciplines often share similar problems but have attempted different solutions. More interdisciplinary collaboration and cross-fertilization of ideas may help to ensure that the most effective strategies are widely shared.
7. Widespread concerns about the veracity of the scientific literature can be unsettling, but this is an exciting time to be a scientist. The scientific method is still the best route to finding truths about nature and we can leverage scientific methods to study science itself. At this critical juncture, meta-research has a crucial role to play in guiding scientists' attempts to calibrate the scientific ecosystem toward higher standards of efficiency, quality, and credibility.

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.

ACKNOWLEDGMENTS

The Meta-Research Innovation Center at Stanford (METRICS) is supported by a grant from the Laura and John Arnold Foundation. The Meta-Research Innovation Center Berlin (METRIC-B)

is supported by a grant from the Einstein Foundation and Stiftung Charitè (Einstein BIH Visiting Fellowship).

LITERATURE CITED

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. 2011. Public availability of published research data in high-impact journals. *PLOS ONE* 6(9):e24357

Altman DG. 1994. The scandal of poor medical research. *BMJ* 308(6924):283–84

Altman DG, Simera I. 2016. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. *J. R. Soc. Med.* 109(2):67–77

Anderson MS, Ronning EA, Devries R, Martinson BC. 2010. Extending the Mertonian norms: scientists' subscription to norms of research. *J. High. Educ.* 81(3):366–93

Armitage P, McPherson CK, Rowe BC. 1969. Repeated significance tests on accumulating data. *J. R. Stat. Soc. Ser. A.* 132(2):235–44

Avey MT, Moher D, Sullivan KJ, Fergusson D, Griffin G, et al. 2016. The devil is in the details: incomplete reporting in preclinical animal research. *PLOS ONE* 11(11):e0166733

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452–54

Bakker M, van Dijk A, Wicherts JM. 2012. The rules of the game called psychological science. *Perspect. Psychol. Sci.* 7(6):543–54

Bakker M, Wicherts JM. 2011. The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43(3):666–78

Bakker M, Wicherts JM. 2014. Outlier removal, sum scores, and the inflation of the type I error rate in independent samples *t* tests: the power of alternatives and recommendations. *Psychol. Methods* 19(3):409–27

Barba LA. 2018. Terminologies for reproducible research. arXiv:1802.03311 [cs.DL]

Begley CG, Ellis LM. 2012. Raise standards for preclinical cancer research. *Nature* 483(7391):531–33

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, et al. 2017. Redefine statistical significance. *Nat. Hum. Behav.* 2(1):6–10

Bennett CM, Miller MB, Wolford GL. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. *NeuroImage* 47(S1):S39–41

Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, et al. 2012. Design and update of a classification system: the UCSD map of science. *PLOS ONE* 7(7):e39464

Boutron I, Dutton S, Ravaud P, Altman DG. 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 303(20):2058–64

Brown NJL, Heathers JAJ. 2017. The GRIM Test: a simple technique detects numerous anomalies in the reporting of results in psychology. *Soc. Psychol. Person. Sci.* 8(4):363–69

Bruce R, Chauvin A, Trinquart L, Ravuad P, Boutron I. 2016. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Med.* 14:85

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–36

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, et al. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2(9):637–44

Carlisle JB. 2012. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia* 67(5):521–37

Carp J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* 63(1):289–300

Chalmers I, Glasziou P. 2009. Avoidable waste in the production and reporting of research evidence. *Lancet* 374(9683):86–89

Chambers CD. 2013. *Registered Reports*: a new publishing initiative at *Cortex*. *Cortex* 49(3):609–10

Chambers CD, Mellor DT. 2018. Protocol transparency is vital for registered reports. *Nat. Hum. Behav.* 2:791–92

Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 291(20):2457–65

Chavalarias D, Ioannidis JPA. 2010. Science mapping analysis characterizes 235 biases in biomedical research. *J. Clin. Epidemiol.* 63(11):1205–15

Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. 2016. Evolution of reporting *P* values in the biomedical literature, 1990–2015. *JAMA* 315(11):1141–48

Ciani O, Buyse M, Garside R, Pavey T, Stein K, et al. 2013. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. *BMJ* 346:f457

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65:145–53

Counsell A, Harlow LL. 2017. Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Can. Psychol./Psychol. Can.* 58(2):140–47

Cramer AOJ, van Ravenzwaaij D, Matzke D, Steingroever H, Wetzels R, et al. 2016. Hidden multiplicity in exploratory multiway ANOVA: prevalence and remedies. *Psychon. Bull. Rev.* 23(2):640–47

Cumming G, Fidler F, Leonard M, Kalinowski P, Christiansen A, et al. 2007. Statistical reform in psychology: Is anything changing? *Psychol. Sci.* 18(3):230–32

Dal-Ré R, Ioannidis JPA, Bracken MB, Buffler PA, Chan A-W, et al. 2014. Making prospective registration of observational research a reality. *Sci. Transl. Med.* 6(224):224cm1

Dechartres A, Boutron I, Trinquart L, Charles P, Ravaud P. 2011. Single-center trials show larger treatment effects than multicenter trials: evidence from a meta-epidemiologic study. *Ann. Intern. Med.* 155(1):39–51

Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, et al. 2017. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ* 357:j2490

Dickersin K, Min YI, Meinert CL. 1992. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 267(3):374–78

Dickersin K, Rennie D. 2012. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA* 307(17):1861–64

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, et al. 2015. Using prediction markets to estimate the reproducibility of scientific research. *PNAS* 112(50):15343–47

Dwan K, Altman DG, Clarke M, Gamble C, Higgins JPT, et al. 2014. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLOS Med.* 11(6):e1001666

Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR. 2011. Comparison of protocols and registry entries to published reports for randomised controlled trials. *Cochrane Database Syst. Rev.* **https://doi.org//10.1002/14651858.MR000031.pub2**

Dwan K, Gamble C, Williamson PR, Kirkham JJ, Report. Bias Group. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLOS ONE* 8(7):e66844

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. 1991. Publication bias in clinical research. *Lancet* 337(8746):867–72

Elms AC. 1975. The crisis of confidence in social psychology. *Am. Psychol.* 30(10):967–76

Ernst AF, Albers CJ. 2017. Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ* 5:e3323

Etz A, Vandekerckhove J. 2016. A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE* 11(2):e0149794

Fanelli D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE* 4(5):e5738

Fanelli D. 2010. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE* 4:e10271

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

Fanelli D, Costas R, Ioannidis JPA. 2017. Meta-assessment of bias in science. *PNAS* 114(14):3714–19

Faust D, Meehl PE. 2002. Using meta-scientific studies to clarify or resolve questions in the philosophy and history of science. *Philos. Sci.* 69(S3):S185–96

Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* 20(5):1539–44

Fiedler K. 2011. Voodoo correlations are everywhere—not only in neuroscience. *Perspect. Psychol. Sci.* 6(2):163–71

Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, et al. 2018. Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psych.* In press. **https://doi.org/10.1016/j.joep.2018.10.009**

Franco A, Malhotra N, Simonovits G. 2014. Publication bias in the social sciences: unlocking the file drawer. *Science* 345(6203):1502–5

Franco A, Malhotra N, Simonovits G. 2016. Underreporting in psychology experiments. *Soc. Psychol. Personal. Sci.* 7(1):8–12

Franzoni C, Scellato G, Stephan P. 2011. Changing incentives to publish. *Science* 333(6043):702–3

Fricker RD, Burke K, Han X, Woodall WH. 2019. Assessing the statistical analyses used in *Basic and Applied Social Psychology* after their *p*-value ban. *Am. Stat.* 73(1):374–84

Gelman A, Loken E. 2014. The statistical crisis in science: data-dependent analysis, a "garden of forking paths," explains why many statistically significant comparisons don't hold up. *Am. Sci.* 102:460–65

Gelman A, Stern H. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* 60:328–31

Gigerenzer G. 2004. Mindless statistics. *J. Socio-Econ.* 33:587–606

Giner-Sorolla R. 2012. Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspect. Psychol. Sci.* 7(6):562–71

Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, et al. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 383(9913):267–76

Goldacre B, Drysdale H, Dale A, Milosevic I, Slade E, et al. 2019. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials* 20:118

Goodman SN. 1993. *p* Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137(5):485–96

Goodman SN. 2018. How sure are you of your result? Put a number on it. *Nature* 564:7

Goodman SN. 2019. Why is getting rid of *p*-values so hard? Musings on science and statistics. *Am. Stat.* 73(S1):26–30

Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8(341):1–6

Gopal AD, Wallach JD, Aminawung JA, Gonsalves G, Dal-Ré R, et al. 2018. Adherence to the International Committee of Medical Journal Editors' (ICMJE) prospective registration policy and implications for outcome integrity: a cross-sectional analysis of trials published in high-impact specialty society journals. *Trials* 19(1):448

Grimes DR, Bauch CT, Ioannidis JPA. 2018. Modelling science trustworthiness under publish or perish pressure. *R. Soc. Open Sci.* 5(1):171511

Hardwicke TE, Frank MC, Vazire S, Goodman SN. 2019a. Should psychology journals adopt specialized statistical review? *Adv. Methods Pract. Psychol. Sci.* **https://doi.org/10.1177/2515245919858428**

Hardwicke TE, Ioannidis JPA. 2018a. Mapping the universe of registered reports. *Nat. Hum. Behav.* 2:793–96

Hardwicke TE, Ioannidis JPA. 2018b. Populating the Data Ark: an attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE* 13(8):e0201856

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(8):180448

Hardwicke TE, Wallach JD, Kidwell MC, Ioannidis JPA. 2019b. An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). MetaArXiv, Apr. 28. **https://doi.org/10.31222/osf.io/6uhg5**

Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. 2016. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 352:i493

Higginson AD, Munafò MR. 2016. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLOS Biol*. 14(11):e2000995

Hoekstra R, Finch S, Kiers HAL, Johnson A. 2006. Probability as certainty: dichotomous thinking and the misuse of *p* values. *Psychon. Bull. Rev.* 13(6):1033–37

IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. 2015. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *J. Clin. Epidemiol.* 68(8):860–69

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med*. 2(8):e124

Ioannidis JPA. 2008. Why most discovered true associations are inflated. *Epidemiology* 19(5):640–48

Ioannidis JPA. 2012. Why science is not necessarily self-correcting. *Perspect. Psychol. Sci.* 7(6):645–54

Ioannidis JPA. 2014. How to make more published research true. *PLOS Med*. 11(10):e1001747

Ioannidis JPA. 2015. Handling the fragile vase of scientific practices. *Addiction* 110(1):9–10

Ioannidis JPA. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q*. 94(3):485–514

Ioannidis JPA. 2017. The reproducibility wars: successful, unsuccessful, uninterpretable, exact, conceptual, triangulated, contested replication. *Clin. Chem.* 63(5):943–45

Ioannidis JPA. 2018a. Meta-research: why research on research matters. *PLOS Biol*. 16(3):e2005468

Ioannidis JPA. 2018b. The proposal to lower *P* value thresholds to .005. *JAMA* 319(14):1429–30

Ioannidis JPA. 2019. Retiring statistical significance would give bias a free pass. *Nature* 567:461

Ioannidis JPA, Caplan AL, Dal-Ré R. 2017a. Outcome reporting bias in clinical trials: why monitoring matters. *BMJ* 356:j408

Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. 2015. Meta-research: evaluation and improvement of research methods and practices. *PLOS Biol*. 13(10):e1002264

Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, et al. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383(9912):P166–75

Ioannidis JPA, Stanley TD, Doucouliagos H. 2017b. The power of bias in economics research. *Econ. J.* 127(605):F236–65

Ioannidis JPA, Trikalinos TA. 2007. An exploratory test for an excess of significant findings. *Clin. Trials.* 4(3):245–53

Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. 2016. Reproducible research practices and transparency across the biomedical literature. *PLOS Biol*. 14(1):e1002333

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23(5):524–32

Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. 2015. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Med*. 13:282

Justice AC, Cho MK, Winker MA, Berlin JA, Rennie D. 1998. Does masking author identity improve peer review quality? A randomized controlled trial. *JAMA* 280(3):240–42

Kaplan RM, Irvin VL. 2015. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLOS ONE* 10(8):e0132382

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* 2(3):196–217

Kimmelman J, Mogil JS, Dirnagl U. 2014. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLOS Biol*. 12(5):e1001863

Klein O, Hardwicke TE, Aust F, Breuer J, Danielsson H, et al. 2018. A practical guide for transparency in psychological science. *Collab. Psychol.* 4(1):20

Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, et al. 2018. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1(4):443–90

Lane DM, Sándor A. 2009. Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychol. Methods* 14(3):239–57

Lash TL, Vandenbroucke JP. 2012. Should preregistration of epidemiologic study protocols become compulsory? *Epidemiology* 23(2):184–88

Lazic SE, Clarke-Williams CJ, Munafò MR. 2018. What exactly is '*N*' in cell culture and animal experiments? *PLOS Biol.* 16(4):e2005282

Leamer EE. 1983. Let's take the con out of econometrics. *Am. Econ. Rev.* 73(1):31–43

Liddell T, Kruschke JK. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *J. Exp. Soc. Psychol.* 79:328–48

Loder E, Loder S, Cook S. 2018. Characteristics and publication fate of unregistered and retrospectively registered clinical trials submitted to the *BMJ* over 4 years. *BMJ Open* 8(2):e020037

Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage JM, et al. 2014. A systematic statistical approach to evaluating evidence from observational studies. *Annu. Rev. Stat. Appl.* 1:11–39

Makel MC, Plucker JA, Hegarty B. 2012. Replications in psychology research. *Perspect. Psychol. Sci.* 7:537–42

Marwick B, Boettiger C, Mullen L. 2017. Packaging data analytical work reproducibly using R (and friends). *Am. Stat.* 72(1):80–88

Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. 2009. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA* 302(9):977–84

Matzke D, Nieuwenhuis S, van Rijn H, Slagter HA, van der Molen MW, Wagenmakers EJ. 2015. The effect of horizontal eye movements on free recall: a preregistered adversarial collaboration. *J. Exp. Psychol. Gen.* 144(1):e1–15

Mayo DG. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge, UK: Cambridge Univ. Press

McGillivray B, De Ranieri E. 2018. Uptake and outcome of manuscripts in *Nature* journals by review model and author characteristics. *Res. Integr. Peer Rev.* 3:5

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. 2019. Abandon statistical significance. *Am. Stat.* 73(1):235–45

Merton RK. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: Univ. Chicago Press

Meyer MN. 2018. Practical tips for ethical data sharing. *Adv. Methods Pract. Psychol. Sci.* 1(1):131–44

Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, et al. 2014. Promoting transparency in social science research. *Science* 343(6166):30–31

Moher D, Dulberg CS, Wells GA. 1994. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 272(2):122–24

Moher D, Jones A, Lepage L, CONSORT (Consol. Stand. Rep. Trials) Group. 2001. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 285(15):1992–95

Moher D, Naudet F, Cristea IA, Miedema F, Ioannidis JPA, Goodman SN. 2018. Assessing scientists for hiring, promotion, and tenure. *PLOS Biol.* 16(3):e2004089

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(1):150547

Moshontz H, Campbell L, Ebersole CR, IJzerman H, Urry HL, et al. 2018. The Psychological Science Accelerator: advancing psychology through a distributed collaborative network. *Adv. Methods Pract. Psychol. Sci.* 1(4):501–15

Mulkay MJ. 1976. Norms and ideology in science. *Soc. Sci. Inf.* 15(4–5):637–56

Munafò MR, Nosek BA, Bishop BVM, Button KS, Chambers CD, et al. 2017. A manifesto for reproducible science. *Nat. Hum. Behav.* 1:0021

Murad MH, Wang Z. 2017. Guidelines for reporting meta-epidemiological methodology research. *Evid. Based Med.* 22(4):139–42

Naudet F, Sakarovitch C, Janiaud P, Cristea IA, Fanelli D, et al. 2018. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in the *BMJ* and *PLOS Medicine*. *BMJ* 360:k400

Nelson LD, Simmons J, Simonsohn U. 2018. Psychology's renaissance. *Annu. Rev. Psychol.* 69(1):511–34
Nieuwenhuis S, Forstmann BU, Wagenmakers E-J. 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14(9):1105–7
Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, et al. 2015. Promoting an open research culture. *Science* 348(6242):1422–25
Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *PNAS* 115(11):2600–6
Nosek BA, Errington TM. 2017. Making sense of replications. *eLife* 6:e23383
Nosek BA, Spies JR, Motyl M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7(6):615–31
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. *Collab. Psychol.* 3(1):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(4):1205–26
O'Boyle EH, Banks GC, Gonzalez-Mulé E. 2013. The Chrysalis Effect: how ugly data metamorphosize into beautiful articles. *Acad. Man. Proc.* 43(2):376–99
Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
Orben A, Przybylski AK. 2019. The association between adolescent well-being and digital technology use. *Nat. Hum. Behav.* 3:173–82
Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. 2016. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. *PLOS ONE* 11(7):e0159267
Pashler H, Wagenmakers EJ. 2012. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7(6):528–30
Patel CJ, Burford B, Ioannidis JPA. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* 68(9):1046–58
Pereira TV, Horwitz RI, Ioannidis JPA. 2012. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 308(16):1676–84
Phillips CV. 2004. Publication bias in situ. *BMC Med. Res. Methods* 4:20
Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, et al. 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* 18(2):115–26
Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10(9):712
Robinson KA, Goodman SN. 2011. A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Ann. Intern. Med.* 154(1):50–55
Rosenthal R. 1966. *Experimenter Effects in Behavioral Research*. East Norwalk, CT: Appleton-Century-Crofts
Rosenthal R. 1979. The file drawer problem and tolerance for null results. *Psychol. Bull.* 86(3):638–41
Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. 2012. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 344:d7292
Ross-Hellauer T, Görögh E. 2019. Guidelines for open peer review implementation. *Res. Integr. Peer Rev.* 4:4
Rowhani-Farid A, Barnett AG. 2016. Has open data arrived at the *British Medical Journal* (*BMJ*)? An observational study. *BMJ Open* 6(10):e011784
Sargent RM. 1999. *Francis Bacon: Selected Philosophical Works*. Indianapolis, IN: Hackett
Schatz P, Jay KA, McComb J, McLaughlin JR. 2005. Misuse of statistical tests in archives of clinical neuropsychology publications. *Arch. Clin. Neuropsych.* 20(8):1053–59
Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. 2014. Interpreting observational studies: why empirical calibration is needed to correct *p*-values. *Stat. Med.* 33(2):209–18
Scott A, Rucklidge JJ, Mulder RT. 2015. Is mandatory prospective trial registration working to prevent publication of unregistered trials and selective outcome reporting? An observational study of five psychiatry journals that mandate prospective clinical trial registration. *PLOS ONE* 10(8):e0133718
Sedlmeier P, Gigerenzer G. 1989. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105(2):309–16

Seminara D, Khoury MJ, O'Brien TR, Manolio T, Gwinn ML, et al. 2007. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* 18(1):1–8

Shamseer L, Hopewell S, Altman DG, Moher D, Schulz KF. 2016. Update on the endorsement of CONSORT by high impact factor journals: a survey of journal "Instructions to Authors" in 2014. *Trials* 17(1):301

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(3):337–56

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Simonsohn U. 2013. Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychol. Sci.* 24(10):1875–88

Smaldino PE, McElreath R. 2016. The natural selection of bad science. *R. Soc. Open Sci.* 3(9):160384

Spellman BA. 2015. A short (personal) future history of Revolution 2.0. *Personal. Psych. Sci.* 10(6):886–99

Stanley DJ, Spence JR. 2014. Expectations for replications: Are yours realistic? *Perspect. Psychol. Sci.* 9(3):305–18

Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. 2016. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* 11(5):702–12

Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* 54(285):30–34

Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *PNAS* 115(11):2584–89

Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. 2007. The use of statistics in medical research. *Am. Stat.* 61(1):47–55

Szucs D, Ioannidis JPA. 2017a. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biol.* 15(3):e2000797

Szucs D, Ioannidis JPA. 2017b. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front. Hum. Neurosci.* 11:390

Tierney JF, Vale C, Riley R, Smith CT, Stewart L, et al. 2015. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLOS Med.* 12(7):e1001855

Trinquart L, Dunn AG, Bourgeois FT. 2018. Registration of published randomized trials: a systematic review and meta-analysis. *BMC Med.* 16(1):173

Tsakiris M, Martin R, Wagemans J. 2018. Re-thinking *Cognition*'s open data policy: responding to Hardwicke and colleagues' evaluation of its impact. *Cognition*. **https://doi.org/10.1016/j.cognition.2018.10.008**

Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. 2012. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst. Rev.* **https://doi.org/10.1186/2046-4053-1-60**

van Dalen HP, Henkens K. 2012. Intended and unintended consequences of a publish-or-perish culture: a worldwide survey. *J. Am. Soc. Inf. Sci. Tech.* 63(7):1282–93

van Dongen NNN, van Doorn J, Gronau QF, van Ravenzwaaij D, Hoekstra R, et al. 2019. Multiple perspectives on inference for two simple statistical scenarios. *Am. Stat.* 73(S1):328–39

Vanpaemel W, Vermorgen M, Deriemaecker L, Storms G. 2015. Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra* 1(1):1–5

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, et al. 2013. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* 1:e148

Vazire S. 2017. Quality uncertainty erodes trust in science. *Collab. Psychol.* 13(4):411–17

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, et al. 2014. The availability of research data declines rapidly with article age. *Curr. Biol.* 24(1):94–97

Voytek B. 2016. The virtuous cycle of a data ecosystem. *PLOS Comput. Biol.* 12(8):e1005037

Vul E, Harris C, Winkielman P, Pashler H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4(3):274–90

Wagenmakers E-J. 2007. A practical solution to the pervasive problems of *p* values. *Psychon. Bull. Rev.* 14(5):779–804

Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas JLJ, Kievit RA. 2012. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* 7(6):632–38

Walker RL, Sykes L, Hemmelgarn BR, Quan H. 2010. Authors' opinions on publication in relation to annual performance assessment. *BMC Med. Educ.* 10:21

Wallach JD, Boyack KW, Ioannidis JPA. 2018. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biol.* 16(11):e2006930

Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JPA. 2017. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern. Med.* 177(4):554–60

Wasserstein RL, Schirm AL, Lazar NA. 2019. Moving to a world beyond "$p < 0.05$." *Am. Stat.* 73(S1):1–19

Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. 2018. Why we need to report more than "data were analyzed by *t*-tests or ANOVA." *eLife* 7:e36163

Weissgerber TL, Milic NM, Winham SJ, Garovic VD. 2015. Beyond bar and line graphs: time for a new data presentation paradigm. *PLOS Biol.* 13(4):e1002128

Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. *Am. Psychol.* 61(7):726–28

Wiseman R, Watt C, Kornbrot D. 2019. Registered reports: an early example and analysis. *PeerJ* 7:e6232

World Med. Assoc. 2013. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 310(20):2191–94

Wu L, Wang D, Evans JA. 2019. Large teams develop and small teams disrupt science and technology. *Nature* 566(7744):378–82

Yong E. 2012. Replication studies: bad copy. *Nature* 485(7398):298–300

Young NS, Ioannidis JPA, Al-Ubaydli O. 2008. Why current publication practices may distort science. *PLOS Med.* 5(10):e201

Zarin DA, Tse T, Ide C. 2005. Trial registration at ClinicalTrials.gov between May and October 2005. *New Engl. J. Med.* 353(26):2779–87