ANNUAL REVIEWS

Annual Review of Psychology

Replicability, Robustness, and Reproducibility in Psychological Science

Brian A. Nosek,[1,2] Tom E. Hardwicke,[3] Hannah Moshontz,[4] Aurélien Allard,[5] Katherine S. Corker,[6] Anna Dreber,[7] Fiona Fidler,[8] Joe Hilgard,[9] Melissa Kline Struhl,[2] Michèle B. Nuijten,[10] Julia M. Rohrer,[11] Felipe Romero,[12] Anne M. Scheel,[13] Laura D. Scherer,[14] Felix D. Schönbrodt,[15] and Simine Vazire[16]

[1]Department of Psychology, University of Virginia, Charlottesville, Virginia 22904, USA; email: ban2b@virginia.edu

[2]Center for Open Science, Charlottesville, Virginia 22903, USA

[3]Department of Psychology, University of Amsterdam, 1012 ZA Amsterdam, The Netherlands

[4]Addiction Research Center, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA

[5]Department of Psychology, University of California, Davis, California 95616, USA

[6]Psychology Department, Grand Valley State University, Allendale, Michigan 49401, USA

[7]Department of Economics, Stockholm School of Economics, 113 83 Stockholm, Sweden

[8]School of Biosciences, University of Melbourne, Parkville VIC 3010, Australia

[9]Department of Psychology, Illinois State University, Normal, Illinois 61790, USA

[10]Meta-Research Center, Tilburg University, 5037 AB Tilburg, The Netherlands

[11]Department of Psychology, Leipzig University, 04109 Leipzig, Germany

[12]Department of Theoretical Philosophy, University of Groningen, 9712 CP Groningen, The Netherlands

[13]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands

[14]University of Colorado Anschutz Medical Campus, Aurora, Colorado 80045, USA

[15]Department of Psychology, Ludwig Maximilian University of Munich, 80539 Munich, Germany

[16]School of Psychological Sciences, University of Melbourne, Parkville VIC 3052, Australia

Annu. Rev. Psychol. 2022. 73:719–48

First published as a Review in Advance on October 19, 2021

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-020821-114157

Keywords

replication, reproducibility, robustness, generalizability, research methods, statistical inference, validity, theory, metascience

Abstract

Replication—an important, uncommon, and misunderstood practice—is gaining appreciation in psychology. Achieving replicability is important for making research progress. If findings are not replicable, then prediction and theory development are stifled. If findings are replicable, then interrogation of their meaning and validity can advance knowledge. Assessing replicability can be productive for generating and testing hypotheses by actively confronting current understandings to identify weaknesses and spur innovation. For psychology, the 2010s might be characterized as a decade of active confrontation. Systematic and multi-site replication projects assessed current understandings and observed surprising failures to replicate many published findings. Replication efforts highlighted sociocultural challenges such as disincentives to conduct replications and a tendency to frame replication as a personal attack rather than a healthy scientific practice, and they raised awareness that replication contributes to self-correction. Nevertheless, innovation in doing and understanding replication and its cousins, reproducibility and robustness, has positioned psychology to improve research practices and accelerate progress.

Contents

INTRODUCTION

The 2010s were considered psychology's decade of crisis (Giner-Sorolla 2019, Hughes 2018), revolution (Spellman 2015, Vazire 2018), or renaissance (Nelson et al. 2018), depending on one's perspective. For decades, methodologists had warned about the deleterious impacts of an overemphasis on statistical significance ($p < 0.05$), publication bias, inadequate statistical power, weak specification of theories and analysis plans, and lack of replication of published findings (Cohen 1973, 1994; Greenwald 1975; Meehl 1978; Rosenthal 1979; Sterling 1959). However, those worries had little impact until conceptual and empirical contributions illustrated their potential ramifications for research credibility (Bakker et al. 2012, Open Sci. Collab. 2015, Simmons et al. 2011, Wagenmakers et al. 2011). This evidence catalyzed innovation to assess and improve credibility. Large initiatives produced surprising failures to replicate published evidence, and researchers debated the role and meaning of replication in advancing knowledge. In this review, we focus on the last 10 years of evidence and accumulated understanding of replication and its cousins, robustness and reproducibility.

WHAT ARE REPRODUCIBILITY, ROBUSTNESS, AND REPLICABILITY?

Replication refers to testing the reliability of a prior finding with different data. Robustness refers to testing the reliability of a prior finding using the same data and a different analysis strategy. Reproducibility refers to testing the reliability of a prior finding using the same data and the same analysis strategy (Natl. Acad. Sci. Eng. Med. 2019). Each of the three notions plays an important role in assessing credibility.

Reproducibility

In principle, all reported evidence should be reproducible. If someone applies the same analysis to the same data, the same result should occur. Reproducibility tests can fail for two reasons. A process reproducibility failure occurs when the original analysis cannot be repeated because of the unavailability of data, code, information needed to recreate the code, or necessary software or tools. An outcome reproducibility failure occurs when the reanalysis obtains a different result than the one reported originally. This can occur because of an error in either the original or the reproduction study.

Achieving reproducibility is a basic foundation of credibility, and yet many efforts to test reproducibility reveal success rates below 100%. For example, Artner and colleagues (2020) successfully reproduced just 70% of the 232 findings analyzed, and 18 of those were reproduced only after deviating from the analysis reported in the original papers (see also Bakker & Wicherts 2011; Hardwicke et al. 2018, 2021; Maassen et al. 2020; Nuijten et al. 2016). Whereas an outcome reproducibility failure suggests that the original result may be wrong, a process reproducibility failure merely indicates that the original result cannot be verified. Either reason challenges credibility and increases uncertainty about the value of investing additional resources to replicate or extend the findings (Nuijten et al. 2018). Sharing data and code reduces process reproducibility failures (Kidwell et al. 2016), which can reveal more outcome reproducibility failures (Hardwicke et al. 2018, 2021; Wicherts et al. 2011).

Robustness

Some evidence is robust across reasonable variations in analysis, and some evidence is fragile, meaning that support for the finding is contingent on specific decisions such as which observations are excluded and which covariates are included. For example, Silberzahn and colleagues

(2018) gave 29 analysis teams the same data to answer the same question and observed substantial variation in the results (see also Botvinik-Nezer et al. 2020). A fragile finding is not necessarily wrong, but fragility is a risk factor for replicability and generalizability. Moreover, without precommitment to an analysis plan, a fragile finding can amplify concerns about *p*-hacking and overfitting that reduce credibility (Simonsohn et al. 2020, Steegen et al. 2016).

Replicability

The credibility of a scientific finding depends in part on the replicability of the supporting evidence. For some, replication is even the sine qua non of what makes an empirical finding a scientific finding (see Schmidt 2009 for a review). Given its perceived centrality and its substantial and growing evidence base in psychological science, we devote the remainder of this article to replicability. Replication seems straightforward—do the same study again and see if the same outcome recurs—but it is not easy to determine what counts as the same study or same outcome.

How do we do the study again? There is no such thing as an exact replication. Even the most similar study designs will have inevitable, innumerable differences in sample units, settings, treatments, and outcomes (Shadish et al. 2002). This fact creates a tension: If we can never redo the same study, how can we conduct a replication? One way to resolve the tension is to accept that every study is unique; the evidence it produces applies only to a context that will never occur again (Gergen 1973). This answer is opposed to the idea that science accumulates evidence and develops explanations for generalizable knowledge.

Another way to resolve the tension is to understand replication as a theoretical commitment (Nosek & Errington 2020a, Zwaan et al. 2018): A study is a replication when the innumerable differences from the original study are believed to be irrelevant for obtaining the evidence about the same finding. The operative phrase is "believed to be." Because the replication context is unique, we cannot know with certainty that the replication meets all of the conditions necessary to observe outcomes consistent with the prior evidence. However, our existing theories and understanding of the phenomenon provide a basis for concluding that a study is a replication. The evidence provided by the replication updates our confidence in the replicability of the finding and our understanding of the conditions necessary or sufficient for replicability to occur.

Because every replication is different from every prior study, every replication is a test of generalizability, but the reverse is not true. A generalizability test is a replication only if all outcomes of the test alter our confidence in the original finding (Nosek & Errington 2020a). For example, if positive outcomes increase confidence and expand generalizability, but negative outcomes merely identify a potential boundary condition and do not alter confidence in the original finding, then it is a generalizability test and not a replication. Applying this framework, the term "conceptual replication" has often been used to describe generalizability tests, not replications, because they are interpreted as supporting the interpretation of a finding but rarely as disconfirming the finding.

Replication as a theoretical commitment leads to considering distinctions such as the one between direct and conceptual replication as counterproductive (Machery 2020, Nosek & Errington 2020a). This position guides this review but is not uncontested. For alternative perspectives about the value of replication and for terminological distinctions among types of replications, readers are referred to Crandall & Sherman (2016), LeBel et al. (2018), Schwarz & Strack (2014), Simons (2014), Stroebe & Strack (2014), Wilson et al. (2020), and Zwaan et al. (2018).

How do we decide whether the same outcome occurred? Empirical evidence rarely provides simple answers or definitiveness, but psychological researchers routinely draw dichotomous

conclusions, often based on whether or not they obtain $p < 0.05$, despite persistent exhortations by methodologists. The desire for dichotomous simplicity occurs with replications too, leading researchers to simply ask, Did the study replicate? Some dichotomous thinking is the result of poor reasoning from null hypothesis significance testing. Some dichotomous thinking is also reasonable as a heuristic for efficient communication. Simplified approaches may be sufficient when the tested theories and hypotheses are underdeveloped. For example, many psychological theories only make a directional prediction with no presumption of rank-ordering conditions or effect sizes. A miniscule effect detected in a sample of 1,000,000 may be treated identically to a massive effect detected in a sample of 10.

There are a variety of options for a dichotomous assessment of replication outcomes, each of which provides some perspective and none of which is definitive. These include assessing whether the replication rejects the null hypothesis ($p < 0.05$) in the same direction as the original study (Camerer et al. 2018, Open Sci. Collab. 2015), computing confidence or prediction intervals of the original or replication findings and assessing whether the other estimate is within an interval or not (Open Sci. Collab. 2015, Patil et al. 2016), assessing whether replication results are consistent with an effect size that could have been detected in the original study (Simonsohn 2015), and subjectively assessing whether the findings are similar (Open Sci. Collab. 2015). There are also some approaches that can be used as continuous measures, such as Bayes factors comparing original and replication findings (Etz & Vandekerckhove 2016) and a Bayesian comparison of the null distribution versus the posterior distribution of the original study (Verhagen & Wagenmakers 2014), although these are often translated into a dichotomous decision about whether a replication failed or succeeded.

As psychological theory and evidence mature, rank-ordering, effect sizes, and moderating influences become more relevant and require a more refined incorporation of replication evidence. Each study contains an operationalization of its conceptual variables of interest, an examination of their relations, and an inference about their meaning. More mature evaluations of replication data reduce the emphasis on individual studies and increase the emphasis on effect sizes and cumulative evidence via meta-analysis or related approaches (Mathur & VanderWeele 2020). Meta-analysis in replication research examines the average effect size, the degree of uncertainty, and the evidence for heterogeneity across samples, settings, treatments, and outcomes (Hedges & Schauer 2019, Landy et al. 2020). When heterogeneity is high, it can indicate that moderating influences need to be identified and tested in order to improve theoretical understanding. Meta-analyses, however, are often undermined by publication bias in favor of significant results as well as other threats to the quality of individual studies (Carter et al. 2019, Rothstein et al. 2005, Vosgerau et al. 2019). Averaging across studies that vary in quality and risk of bias, including when meta-analyzing original and replication studies, can lead to a false sense of precision and accuracy.

Ultimately, replication is a central feature of the ongoing dialogue between theory and evidence: The present understanding is based on the cumulative evidence; areas of uncertainty are identified; tests are conducted to examine that uncertainty; and new evidence is added, reinforcing or reorganizing present understanding. Then the cycle repeats.

A Note on Validity

A finding can be reproducible, robust, replicable, and invalid at the same time. Credibility in terms of reproducibility, robustness, and replicability does not guarantee that the treatments worked as intended, the measures assessed the outcomes of interest, or the interpretations correspond with the evidence produced. However, conducting replications can help identify sources of invalidity if those sources of invalidity are present or absent across replications testing a same phenomenon by

happenstance. Deliberate efforts to root out invalidity through replications can also be productive. Suppose, for example, that the original finding is that increasing empathy reduces racial bias. An observer might suspect that the intervention affected more than empathy, potentially undermining validity. A replication could pursue the same evidence but with an alteration that meets the theoretical conditions for increasing empathy without influencing other variables. Stated this way, the ordinariness and centrality of replication for advancing knowledge become clear. Many replications, in practice, are efforts to root out invalidity due either to questionable research practices that undermine the statistical evidence or to questionable validity of the treatments, measures, and other study features used to produce the existing evidence.

THE STATE OF REPLICABILITY OF PSYCHOLOGICAL SCIENCE

Warning signs that replicability in psychology might be lower than expected or desired have been available for decades. Cohen and others (Button et al. 2013; Cohen 1973, 1992; Sedlmeier & Gigerenzer 1992; Szucs & Ioannidis 2017) noted that the median power of published studies is quite low, often below 0.50. This means, assuming that all effects under study are true and accurately estimated, that one would expect less than 50% of published findings to be statistically significant ($p < 0.05$). However, other evidence suggests that 90% or more of primary outcomes are statistically significant (Fanelli 2010, 2012; Sterling 1959; Sterling et al. 1995). Moreover, a meta-analysis of 44 reviews of statistical power observed a mean statistical power of 0.24 to detect a small effect size ($d = 0.20$) with a false-positive rate of $\alpha = 0.05$, and there was no increase in power from the 1960s through the 2010s (Smaldino & McElreath 2016; see also Maxwell 2004). The evidence of low power and high positive result rates cannot be reconciled easily without inferring the influence of publication bias, whereby negative results are ignored (Greenwald 1975, Rosenthal 1979), or questionable research practices that inflate reported effect sizes (John et al. 2012, Simmons et al. 2011). Despite broad recognition of this disjoint, there have been no systematic efforts to test the credibility of the psychological literature until this past decade.

The replicability of psychological research is unlikely to ever be answered with confidence, as the psychological literature is large, changes constantly, and has ill-defined boundaries. Examining a large-enough random selection of studies to make a precise and meaningful estimate exceeds feasibility. However, progress can be made by benchmarking the replicability of samples of findings against the expectations about their credibility. For example, if an individual finding is regularly cited and used as a support for theory, then there exists an implicit or explicit presumption that the finding is replicable. Likewise, testing any sample of studies from the published literature presents an opportunity to evaluate the replicability of those findings against the expectation that published results in general are credible. Any generalization of replicability estimates to studies that were not included in the sample will involve some uncertainty. This uncertainty will increase as the studies become less similar to the replication sample.

People disagree about what degree of replicability should be expected from the published literature (Gordon et al. 2020). To provide some empirical grounding to these discussions, we summarize recent evidence concerning replicability in psychology. We gathered two types of prominent replication studies conducted during the last decade: (*a*) Systematic replications are replication efforts that define a sampling frame and conduct replications of as many studies in the sampling frame as possible to minimize selection biases, and (*b*) multi-site replications are studies that conduct the same replication protocol in a variety of samples and settings to obtain highly precise estimates of effect size and heterogeneity. In **Figure 1**, we pragmatically summarize the outcomes with two popular criteria for assessing replication success: statistical significance in the same direction and comparison of observed effect sizes.

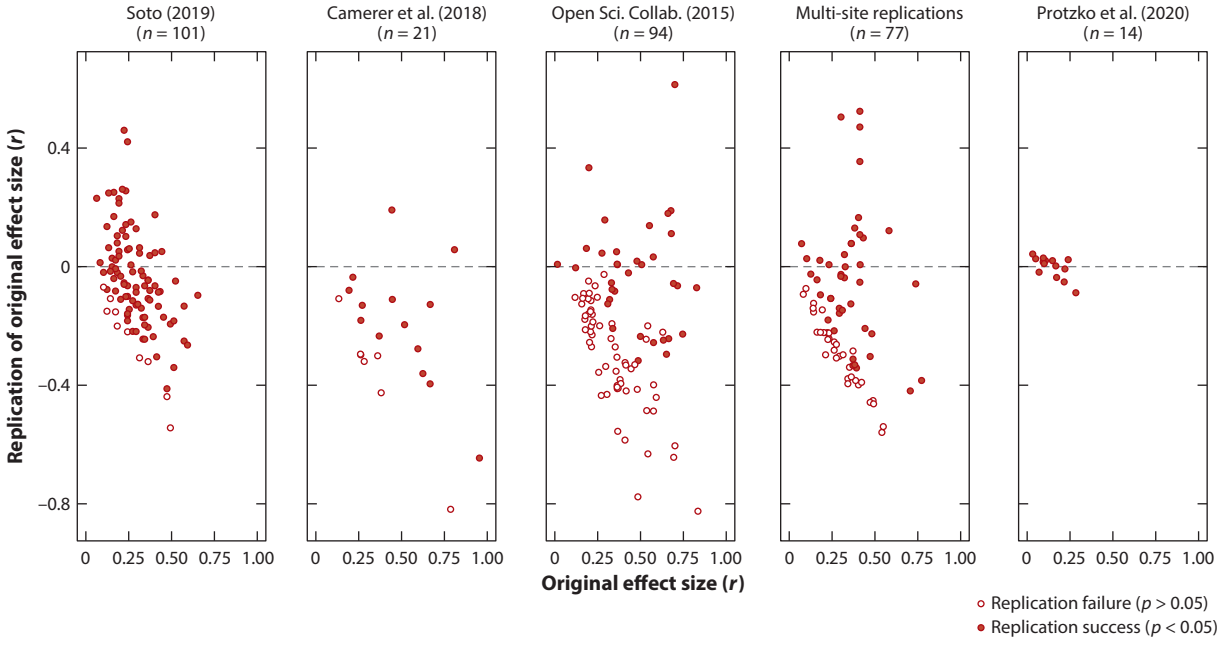

Figure 1

Replication outcomes for three systematic replication studies (Camerer et al. 2018, Open Sci. Collab. 2015, Soto 2019), multi-site replication studies, and a prospective best-practice replication study (Protzko et al. 2020). Values above zero indicate that the replication effect size was larger than the original effect size. Solid circles indicate that replications were statistically significant in the same direction as the original study. Studies with effects that could not be converted to *r* or original studies with null results are excluded.

With regard to systematic replications, Soto (2019) replicated 101 associations between personality traits and outcomes (all measured with self-reports in the replication studies) identified from a published review of the literature and observed that 90% achieved statistical significance in the same direction, with effect sizes 91% as large as those found in the original studies. Camerer and colleagues (2018) replicated 21 social science experiments systematically selected from *Nature* and *Science* articles published between 2010 and 2015; 62% achieved significance in the same direction, with effect sizes 50% as large as in the original studies. Open Sci. Collab. (2015) replicated 100 findings from the articles published in 2008 in three psychology journals; 36% achieved significance in the same direction, with effect sizes 49% as large as in the original studies.

With regard to multi-site replications, these include the series titled "Many Labs" (Ebersole et al. 2016a, 2020; Klein et al. 2014, 2018, 2019); registered replication reports, primarily from the journal *Advances in Methods and Practices in Psychological Science* (Alogna et al. 2014, Bouwmeester et al. 2017, Cheung et al. 2016, Colling et al. 2020, Eerland et al. 2016, Hagger et al. 2016, McCarthy et al. 2018, O'Donnell et al. 2018, Verschuere et al. 2018, Wagenmakers et al. 2016); papers from the Collaborative Replications and Education Project (Ghelfi et al. 2020, Leighton et al. 2018, Wagge et al. 2018); and other similar efforts (Dang et al. 2021; ManyBabies Consort. 2020; McCarthy et al. 2018, 2021; Moran et al. 2020; Schweinsberg et al. 2016). Collectively ($n = 77$), 56% of multi-site replications reported statistically significant evidence in the same direction, with effect sizes 53% as large as in the original studies (**Figure 1**).

Considering all replications ($n = 307$), 64% reported statistically significant evidence in the same direction, with effect sizes 68% as large as in the original studies. Moreover, the sample size used in the replication studies was on average 15.1 times the size used in the original studies (with

a median of 2.8 and a standard deviation of 32.8); this allowed for more precise estimates of effect size and heterogeneity and led to a relatively generous dichotomous definition of "success" as high power to detect a significant effect in the same direction as the original even if the effect size was much smaller in the replication study compared with the original study.

We cannot measure the overestimation of replicability or effect sizes in psychology in general, but we can conclude that replicability challenges are observed almost everywhere that has undergone systematic examination. To preserve the view that the psychological literature is highly replicable, we would need to observe at least some sampling strategies that reveal high replicability and find evidence explaining how these systematic and multi-site replications underestimated replicability.

WHAT REPLICATES AND WHAT DOES NOT?

Some replications produce evidence consistent with the original studies; others do not. Why is that? Knowledge about the correlates and potential causes could help improve interventions to increase replicability. We discuss three overlapping classes of correlates: theoretical maturity, features of the original studies, and features of the replication studies.

Theoretical Maturity

We do not know in advance whether a phenomenon exists, but we might have an estimate of its prior probability based on existing knowledge. A phenomenon predicted by a well-established theory with a strong track record of making successful predictions and withstanding falsification attempts might elicit a high prior probability, whereas a phenomenon predicted by a new theory that has not yet been tested might elicit a low prior probability. Replicability should therefore be related to theoretical maturity (Cronbach & Meehl 1955, Muthukrishna & Henrich 2019).

An important component of a study's theoretical maturity is a good definition of how the theory's variables are causally connected. This helps to generate clear predictions and to identify auxiliary hypotheses and boundary conditions. Theory formalization and transparency should also increase replicability, because an appropriate study design is easier to determine. This minimizes hidden moderators that might qualify whether a phenomenon is observed, which are a consequence of underspecified theories.

Even for well-specified theories, any given study design inevitably includes auxiliary hypotheses that are not covered by theory. Many auxiliary hypotheses are implicit and may not even be made consciously (Duhem 1954). For example, even mature psychological theories might not specify all parameters for the physical climate, presence or absence of disabilities, and cultural context because of seeming obviousness, theoretical complexity, or failure to consider their influence. An insufficient level of detail makes it difficult to identify theoretical expectations and background assumptions that could influence replicability. An original study might observe a true positive and a replication attempt might observe a true negative, even if both studies are well conducted, if our understanding is not yet mature enough to anticipate the consequences of seemingly irrelevant factors in the sample, setting, intervention, and outcome measures (Nosek & Errington 2020a, Stroebe & Strack 2014).

That such contextual sensitivity can occur is a truism, but invoking it to explain the difference between the results of an original study and a replication demands evidence (Simons 2014, Zwaan et al. 2018). Van Bavel and colleagues (2016) observed that judges' ratings of the context sensitivity of the phenomena included in Open Sci. Collab.'s (2015) study were negatively associated with replication success ($r = -0.23$). However, Inbar (2016) observed that the correlation did not hold within social ($r = -0.08$) and cognitive ($r = -0.04$) subdisciplines, which could reflect

confounding by discipline. Appeals to context sensitivity are common in response to failures to replicate (Cesario 2014, Crisp et al. 2014, Dijksterhuis 2018, Ferguson et al. 2014, Gilbert et al. 2016, Schnall 2014, Schwarz & Strack 2014, Shih & Pittinsky 2014). However, there are multiple examples in which presumed context sensitivity, when examined directly, fails to account for replication failures or weaker effect sizes than found in original studies (Ebersole et al. 2016a, 2020; Klein et al. 2014, 2018). Heterogeneity is sometimes observed in replication studies, but it is usually modest and insufficient to make a replicable phenomenon appear or disappear based on factors that would not have been anticipated in advance of conducting the studies (Baribault et al. 2018; Klein et al. 2014, 2018; Olsson-Collentine et al. 2020). Identifying the circumstances in which the replicability of a finding is demonstrated to be contingent on unconsidered factors in the operationalization will advance investigations of the correlates of replicability.

Features of Original Studies

A finding may not be replicable because the original finding is a false positive (or a false negative in the case of the rarely reported null results). If researchers investigate hypotheses with lower prior odds of being true, the false positive rate can be high (Button et al. 2013, Ioannidis 2005). Dreber and colleagues (2015) provided initial evidence that psychologists tend to investigate hypotheses with low prior probabilities. Using Bayesian methods, they derived the median prior odds of a sample of findings replicated by Open Sci. Collab. (2015) to be just 8.8% (with probabilities ranging from 0.7% to 66%). Relatedly, Open Sci. Collab. (2015) reported exploratory evidence that studies with more surprising results were less likely to replicate ($r = -0.24$) (see also Wilson & Wixted 2018).

Original findings based on weak statistical evidence may be more difficult to replicate than original findings based on strong evidence. For example, Open Sci. Collab. (2015) reported the finding from exploratory analyses that lower *p*-values in the original studies were associated with higher likelihood of replication success ($r = -0.33$). In a literature with relatively small sample sizes and a publication bias favoring positive results, large observed effects can be a sign of overestimated or false positives rather than of large true effects (Gelman & Carlin 2014). Studies with larger samples, better measures, and more tightly controlled designs reduce error and produce more credible evidence, along with better insight about what is necessary to observe the effect (Ioannidis 2005, 2014).

Findings reported with low transparency may be difficult to replicate for the mundane reason that it is difficult to understand what was done in the original study. It is rare that the theoretical conditions necessary for observing a finding are well specified and general enough to achieve high replicability without reference to the operationalization of the methods. Errington and colleagues (2021) documented their difficulty in designing 193 cancer biology replications using only the information provided in the original papers and supplements. In no case were they able to create a comprehensive protocol without asking clarifying questions to the original authors. Transparency and sharing of all aspects of the methodology make clear how the original finding was obtained, reduce the burden of making inferences from underspecified theories, and illuminate auxiliary and unstated assumptions about the conditions that are sufficient to observe an original finding. This includes transparent reporting of all analytic decisions and outcomes to avoid unacknowledged "gardens of forking paths" (Gelman & Loken 2013). Making research contents findable, accessible, interoperable, and reusable (FAIR; see Wilkinson et al. 2016), following standards such as the journal article reporting standards (JARS; see Appelbaum et al. 2018), and employing methods such as born-open data sharing (Rouder 2016) can reduce or expose reporting errors, foster more comprehensive reporting of methodology, and clarify the data structure and analytic decisions.

Failing to report transparently the process and the context that produced the original findings may reduce their replicability. For example, researchers are more likely to publish positive findings than negative findings (Franco et al. 2014, 2016; Greenwald 1975). Reporting biases favoring positive results will produce exaggerated effect sizes and false positive rates, particularly in low-powered research contexts (Button et al. 2013; Ioannidis 2005, 2008). Conducting multiple studies and reporting only a subset of results that achieve statistical significance will inevitably reduce replicability if the research includes any false hypotheses (Nuijten et al. 2015, Schimmack 2012). Furthermore, original findings that result from selective reporting, *p*-hacking, or other behaviors that leverage random chance to amplify effect sizes, obtain statistical significance, or specify overfitting models should be less likely to replicate than others (Bakker et al. 2012, Götz et al. 2020, Nelson et al. 2018, Simmons et al. 2011).

Preregistration ensures that all studies are, in principle, discoverable even if they are not reported in published articles. Increasing the transparency and discoverability of all studies will improve the accuracy of meta-analyses and the estimation of likelihood to replicate. Preregistration of analysis plans helps to calibrate confidence on the reliability of unplanned analyses (Nosek et al. 2018, Wagenmakers et al. 2012).

Features of Replication Studies

A finding may not replicate because the replication is a false negative (or a false positive for the relatively rare published null findings). Many of the features that tend to decrease the replicability of an original finding apply to replications, too: small samples, poorly controlled designs, and other factors that reduce statistical power and increase uncertainty (Maxwell et al. 2015). As in the case of original studies, incomplete reporting can also distort the evidence: The probability of a false negative may be exacerbated if replications are subject to a reverse publication bias whereby negative results are more likely to be reported than positive results (Ioannidis & Trikalinos 2005).

Just like original studies, replication attempts can fail due to errors or oversights by the researchers. Blaming a given failure to replicate on researcher incompetence is a hypothesis that requires empirical evidence. So far, in a decade of intense attention to replications and the skills of replicators, there have been frequent assertions (Baumeister 2016, Baumeister & Vohs 2016, Gilbert et al. 2016, Schnall 2014, Schwarz & Strack 2014, Shih & Pittinsky 2014) and little evidence that failures to replicate are due to shortcomings of the replication studies. Virtually all of the replication studies reviewed in **Figure 1** include preregistration of study designs and analysis plans and open sharing of materials and data to facilitate such investigations.

Further, specific cases that have been empirically examined do not support the conclusion that the replication failures occurred because of incompetence or failings to implement and conduct the study appropriately. Gilbert and colleagues (2016) speculated that Open Sci. Collab. (2015) failed to replicate many studies because the replication teams did not conduct the experiments with sufficient fidelity to the originals, and they brought the original authors' lack of endorsement of some protocols as supporting evidence (but see Anderson et al. 2016, Nosek & Gilbert 2017). Both Gilbert and colleagues (2016) and Wilson and colleagues (2020) suggested that a substantial portion of failures to replicate were due to underpowered replications. Ebersole and colleagues (2020) tested these claims by replicating 10 of the replications with extremely high-powered tests (median $N = 1{,}279$), using the same replication protocols in one condition and protocols revised following formal peer review by experts in the other condition. Formal expert review produced little to no increase in effect sizes compared to the original replication protocols (see also Ebersole et al. 2017, Klein et al. 2019).

Predicting Replicability

Given that not all findings are replicable, it would be helpful if we could anticipate replication likelihood in advance of conducting laborious replications. If replicability is predictable, then there is information about credibility in some features of the original studies and findings. This consideration might also foster the development of faster and cheaper indicators of replicability so as to guide attention and resources toward findings that are most valuable to replicate (Isager et al. 2020).

Evidence from three approaches engaging human judgment—surveys, structured elicitation protocols, and prediction markets—suggests that replication outcomes are predictable. Surveys present brief descriptions of the original studies and findings and then average individual estimates about the likelihood of successful replication. Structured elicitations engage small groups of individuals who make private initial judgments and then discuss their estimates and share information among them before making a second final private judgment (Hanea et al. 2017). Structured protocols such as Investigate Discuss Estimate Aggregate (IDEA) rely on the mathematical aggregation of group members' final judgments rather than forcing behavioral consensus in the way traditional Delphi groups do. Prediction markets have participants bet on whether the studies will replicate or not by buying and selling contracts for replications. Contracts representing each study are worth \$1 if the study replicates and \$0 if it does not replicate. The price of a contract is interpreted as the probability that the market predicts the replication outcome to be successful (Dreber et al. 2015).

In all approaches, aggregating across projects, predictions were positively correlated with the observed replication success (with r values of 0.52 for prediction markets, 0.48 for surveys, and 0.75 for elicitations) (**Figure 2**). Using a dichotomous criterion whereby prices above 50 anticipate replication success and prices below 50 anticipate replication failure in the prediction markets, 88 of 123 (72%) findings and 79 of 123 (64%) findings in similar survey formats were predicted accurately. So far, prediction markets on replication outcomes based on effect size have not been very successful (Forsell et al. 2019), whereas survey evidence suggests some success (Landy et al. 2020). Using a survey method, Hoogeveen and colleagues (2020) observed that, for a subset of 27 of 123 prediction-replication pairs, lay people predicted replication success with 59% accuracy, which increased to 67% when they also received information about the strength of the original evidence. There are not yet any studies to assess whether social-behavioral expertise confers any advantage when predicting replication outcomes.

Humans, regardless of expertise, may not be needed at all. Two studies used machine learning models to predict replicability by training predictive models and then doing out-of-sample tests (Altmejd et al. 2019, Yang et al. 2020). The results reinforce the conclusion that statistical properties like the sample sizes, p-values, and effect sizes of the original studies, as well as whether the effects are main effects or interaction effects, are predictive of successful replication (Altmejd et al. 2019). Models trained on the original papers' narrative text performed better than those on reported statistics (Yang et al. 2020). In both studies, the models performed similarly to the prediction markets on the same data. If these findings are themselves replicable, then machine learning algorithms could provide a high-scalable early assessment of replicability and credibility that may inform evaluation, drive resource allocation, and help identify gaps and strengths in the empirical evidence. A third study used a different type of forecasting approach by using the original studies' information and the replication studies' sample size (Pawel & Held 2020). For the comparable samples, the forecasts from the tested statistical methods performed equally well as, or worse than, the prediction markets.

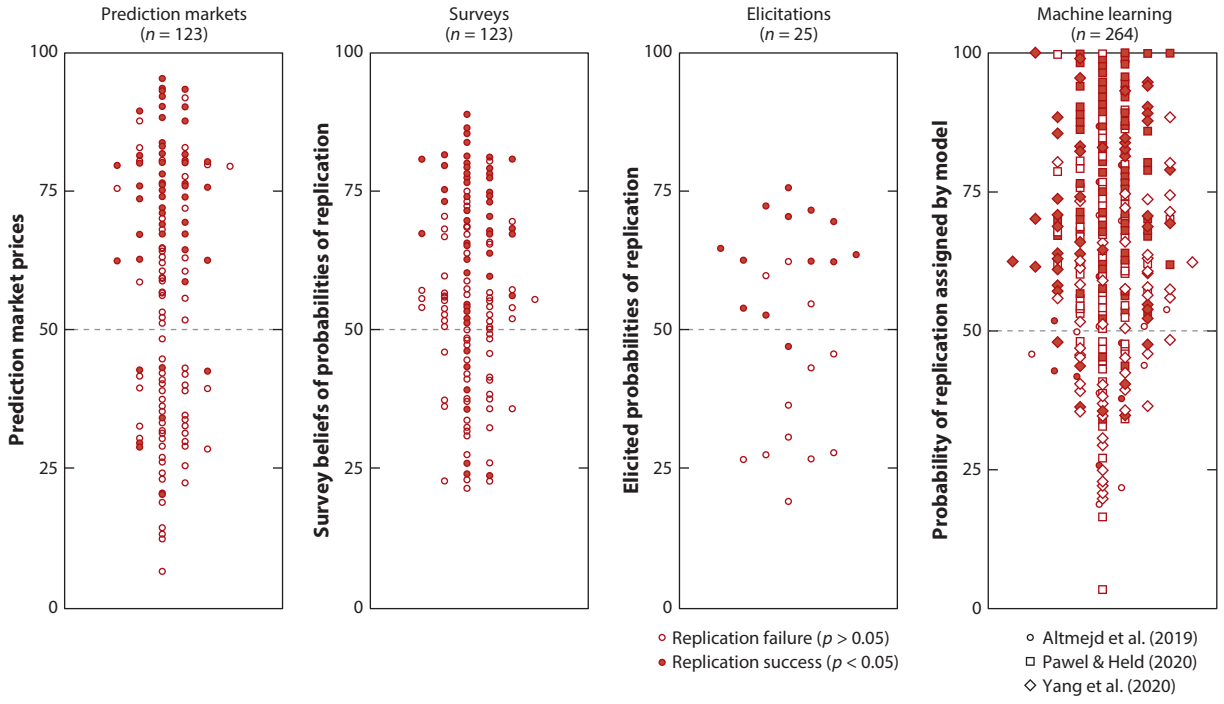

Figure 2

Predictions of replication outcomes across four methods: prediction markets, surveys, elicitations, and machine learning. The figure aggregates 123 prediction-replication pairs for which there are both survey and market predictions and outcomes from five different prediction projects (Camerer et al. 2016, 2018; Dreber et al. 2015; Ebersole et al. 2020; Forsell et al. 2019), 25 elicitations for 25 replications (Wintle et al. 2021), and 264 machine learning scores from three projects (Altmejd et al. 2019, Pawel & Held 2020, Yang et al. 2020). Probabilities of replication were computed on a 0 to 100 scale for all three methods, and all three sets of human predictions were performed by experts.

What Degree of Replicability Should Be Expected?

Non-replicable findings are a risk factor for the progress of research, but it does not follow that a healthy research enterprise is characterized by all findings being replicable (Lewandowsky & Oberauer 2020). It would be possible to achieve near-100% replicability by adopting an extremely conservative research agenda that studies phenomena that are already well understood or have extremely high prior odds. Such an approach would produce nearly zero research progress. Science exists to expand the boundaries of knowledge. In this pursuit, false starts and promising leads that turn out to be dead ends are inevitable. The occurrence of nonreplicability should decline with the maturation of a research topic, but a healthy, theoretically generative research enterprise will include nonreplicable findings. At the same time, a healthy, theoretically generative research enterprise will be constantly striving to improve replicability. Even for the riskiest hypotheses and the earliest ventures into the unknown, design and methodology choices that improve replicability are preferable to those that reduce it.

Improving Replicability

Low replicability is partly a symptom of tolerance for risky predictions and partly a symptom of poor research practices. Persistent low replicability is a symptom of poor research practices.

Replicability can be improved by conducting more severe tests of predictions (Mayo 2018). This involves increasing the strength of methods to amplify signal and reduce error. Increasing the strength of methods entails increasing the number of observations, using stronger measures and manipulations, and improving design with validity checks, piloting, and other validity enhancements (Smith & Little 2018, Vazire et al. 2020). Reducing error entails setting stricter inference criteria (Benjamin et al. 2018, Lakens et al. 2018); guarding against *p*-hacking, hypothesizing after the results are known (or HARKing), and selective reporting by employing preregistration and transparency; and taking alternative explanations seriously by undertaking robustness checks, cross-validation, and internal replications. These improvements are complemented by reporting conclusions that correspond with the evidence presented (Yarkoni 2019), articulating presumed constraints on the generality of the findings (Simons et al. 2017), and calibrating certainty based on the extent to which the statistical inferences could be influenced by prior observation of the data or overfitting (Wagenmakers et al. 2012).

Replicability will be improved if it is easy for anyone to evaluate the severity of the tests and the credibility of the conclusions and to conduct follow-up research. Evaluation is facilitated by maximizing the transparency of the research process, including by sharing methods, materials, procedures, and data; by reporting the timing of decisions and any data dependency in the analysis (Lakens 2019, Nosek et al. 2019); and by making explicit any hidden knowledge that might affect others' evaluation or replication of the research, such as conflicts of interest. Likewise, replication research may increase recognition that effect sizes are overestimated in the published literature and that new research should expect smaller effect sizes and plan for larger samples to detect them (Funder & Ozer 2019, Perugini et al. 2014).

Initial evidence suggests that behavioral changes such as preregistering the study, using large samples, and sharing research materials are associated with high replicability. Protzko and colleagues (2020) implemented these behaviors in a best-practice prospective replication study in which four independent laboratories replicated novel findings in a round-robin format. As shown in **Figure 1**, the large-sample, preregistered original studies elicited relatively small effect sizes compared to the original findings from other replication projects. However, those original effect sizes appear to be credible: Replication effect sizes were 97% as large as in the original studies, suggesting that high replicability is achievable. This study does not, however, provide causal evidence of specific practices that increase replicability.

Structural solutions can improve replicability by incorporating more rigorous research practices into the reward and evaluation systems. For example, Registered Reports is a publishing model in which authors receive in-principle acceptance of their proposed studies based on the importance of the question and the quality of the methodology before the outcomes are known (Chambers 2019). Registered Reports provide a structural solution for selecting good research questions, using appropriately severe methods and procedures, preregistering planned analyses, and presenting work transparently for others to provide critique. Registered Reports are also associated with high rates of sharing of data and materials and higher ratings of quality and rigor compared to regular papers (Soderberg et al. 2021) as well as a much higher proportion of reported null results (Scheel et al. 2020). As of yet, there is no investigation of whether findings from Registered Reports are more replicable on average than other findings. Similarly, encouraging adversarial collaboration could help advance progress, particularly when there are alternative perspectives (Ellemers et al. 2020, Kahneman 2003), and integrating that process with Registered Reports could be particularly fruitful for making commitments and predictions explicit in advance (Nosek & Errington 2020b). Finally, supporting and incentivizing the work of those who find and publicize errors could enhance the replicability of everyone's findings by creating a culture that values "getting it right" rather than simply "getting it published" (Marcus & Oransky 2020).

CULTURAL, SOCIAL, AND INDIVIDUAL CHALLENGES FOR IMPROVING REPLICABILITY

An understanding of the underlying causes of non-replicability and of how to address them is not sufficient to improve replicability. The problems of low-powered research and misuse of null hypothesis significance testing, and their solutions, have been understood since before Jacob Cohen started writing about them in the 1960s (Cohen 1962). Indeed, reflecting on the lack of increase in sample size and power 30 years later, he noted, "I have learned, but not easily, that things take time" (Cohen 1990, p. 1311), and "we must finally rely, as have the older sciences, on replication" (Cohen 1994, p. 1002). Psychological science is a complex system shaped by structural constraints, social contexts, and individual knowledge, biases, and motivations (Smaldino & McElreath 2016). Improving replicability is not just about knowing what to do; it is also about addressing the structural, social, and individual factors that diminish the ability and opportunity to do it (Hardwicke et al. 2020a).

Social and Structural Context

Academic science occurs in a complex system of policies, norms, and incentives that shape decisions about which research is funded, which research gets published, and which researchers get jobs and promotions. This system of policies, norms, and incentives is strong enough that researchers may value behaviors that improve replicability and may know how to perform them, but they may still not do it because the behaviors are not rewarded or are even costly to one's career advancement.

The currency of advancement in academic science is publication (Bakker et al. 2012), and not everything gets published. Positive, novel, tidy results are more likely to get published than negative, replication, or messy results (Giner-Sorolla 2012, Nosek et al. 2012, Romero 2017). Having substantial discretion on which studies and analyses they report, researchers have both the motivation and the opportunity to engage intentionally or unintentionally in behaviors that improve publishability at the cost of credibility.

These trends affect all areas of science, but they might be particularly acute in fields in which there is a lack of consensus about constructs, definitions, and theories (Leising et al. 2020). With incentives rewarding innovation, it is in the researchers' self-interest to avoid using constructs and theories developed by others, as illustrated by the aphorism "Psychologists treat theories like toothbrushes—no self-respecting person wants to use anyone else's" (Mischel 2008). If even using someone else's theory for novel research is an impediment to one's career advancement, it is no surprise that replications are undervalued.

The emphasis on novelty further discourages researchers from adopting rigor-enhancing practices in their own work. Conducting replications could make novel, positive results go away—and the publication prospects with them. Moreover, Bakker and colleagues (2012) provided modeling evidence that if the goal is to create as many positive results as possible, then it is in the researchers' self-interest to run many underpowered studies rather than fewer well-powered ones (see also Gervais et al. 2015, Tiokhin & Derex 2019). The result of the combination of these factors is explosive: a literature filled with novel contributions selectively reported from many underpowered studies, without replications to assess their credibility or improve their precision. The lack of replication also means that there are no costs (reputational or otherwise) for publishing false positives, thus reinforcing the singular emphasis on novelty.

Smaldino & McElreath (2016) incorporated some of these structural incentives into a dynamic model of the research community conducting research and publishing results, and they found that

"the persistence of poor methods. . .requires no conscious strategizing—no deliberate cheating nor loafing—by scientists, only that publication is a principal factor for career advancement." Moreover, they observed that adding replication studies to the picture is insufficient to alter this dysfunctional process, emphasizing that structural change is also necessary for effective reform to happen.

Imbuing scholarly debates about evidence with negative judgments about the researcher's character and intention interferes with the goal of treating replication as ordinary good practice (Meyer & Chabris 2014, Yong 2012). This creates a fraught social environment that discourages researchers from undertaking a skeptical and critical inquiry into existing findings. This dysfunctional social culture might be sustained by the fact that researchers' academic standing is based on their findings rather than their demonstrated rigor and transparency. If reputation is defined by one's findings, then a failure to replicate functionally becomes an attack on one's reputation.

Optimistically, there is conceptual support for social and structural change. Researchers endorse core values of science such as transparency and self-skepticism (Anderson et al. 2007) and disagree with the cultural devaluation of replication. Faced with a choice between a researcher who conducts boring but reproducible research and one who conducts exciting but not reproducible research, raters consistently and strongly favor the former (Ebersole et al. 2016b). Moreover, researchers who take seriously an independent failure to replicate their findings by acknowledging it or conducting follow-up research improve their reputation, even more so than researchers whose findings replicate successfully (Ebersole et al. 2016b, Fetterman & Sassenberg 2015). This suggests that the challenges are rooted in the social and structural features of science, not in the minds of the practicing scientists. Highlighting the consequences of some relevant practices, like small sample size, in decision-making contexts may be sufficient to spur shifts in evaluation (Gervais et al. 2015).

Individual Context

Even if the social environment does not explicitly tie researchers' reputations to their findings, people like to be right (Kunda 1990). However, scientific progress is made by identifying where a current understanding is wrong and generating new ideas that might make it less wrong. Therefore, building a successful scientific career involves getting used to being wrong—a lot. Commenters have promoted the value of cultivating mindsets that embrace getting it right over being right and praise intellectual humility more generally (Ebersole et al. 2016b, Leary et al. 2017, Whitcomb et al. 2017). Intellectual humility may be relevant to a variety of reasoning biases that can interfere with the pursuit of truth. These include confirmation bias, whereby researchers might selectively attend to or create conditions that are consistent with their existing positions (Nickerson 1998); hindsight bias, whereby researchers might revise their theoretical predictions about the results of a replication design after observing the outcomes (Christensen-Szalanski & Willham 1991, Kerr 1998); and outcome bias, whereby researchers might evaluate the quality of a replication design based on whether the outcomes are consistent or inconsistent with their desired outcome (Baron & Hershey 1988, Nosek & Errington 2020b). It is not yet clear whether an intentional embrace of intellectual humility is sufficient to overcome the variety of motivated reasoning biases that help to preserve people's sense of understanding, accuracy, and self-esteem (Kunda 1990). As in other contexts for which biases are difficult to detect and overcome, structural solutions such as preregistration and transparency may be needed to mitigate the opportunity for such reasoning biases to affect judgment or to make them more evident when they do occur.

A CHANGING RESEARCH CULTURE

Psychology in 2021 is different from psychology in 2011. Researchers have accumulated a substantial evidence base about replicability and credibility and how to improve them. Grassroots initiatives have shifted norms, advanced training, and promoted structural change. And journal editors, funders, and leaders have adopted new policies and practices to shift incentives and requirements. These activities amount to a decentralized behavioral change strategy that is transforming how research is done, reported, and evaluated.

Strategy

The culture change movement is composed of many stakeholders making independent decisions about whether and how to change their policies and practices to improve replicability. There has also been collaboration and coordination among stakeholders and grassroots initiatives. Finally, there are organizations such as the Society for the Improvement of Psychological Science and the Center for Open Science (COS) that have the mission to promote culture change toward rigor, transparency, and replicability. COS's culture change strategy, based on Rogers's diffusion model (**Figure 3**), is an example of how decentralized actions by many groups are fostering culture change (Nosek 2019a).

The model extends Rogers's (2003) theoretical work on the diffusion of innovations, which describes how new technologies are first used by innovators and early adopters and then gain mainstream acceptance. This model rests on a few key principles: Culture and behavior changes unfold over time, motivations differ across people and situations between the introduction and the mass adoption of new behaviors, and multiple interdependent interventions are necessary to address variations in motivations and to leverage the adoption by some to stimulate adoption by others.

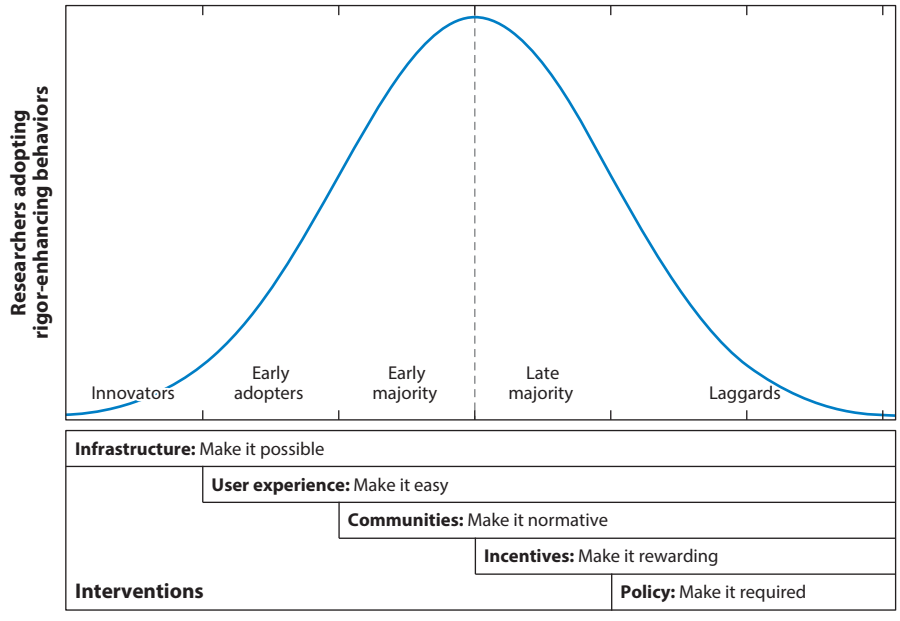

Figure 3

Interdependent interventions for effective culture change extending Rogers's (2003) diffusion model.

According to COS's extension of the diffusion model, for innovators who are motivated to try and test new behaviors, providing infrastructure and tools that make it possible to do so can be sufficient to stimulate adoption. Expanding to early adopters, those motivated by the vision and promise of the new behaviors, requires user-centered attentiveness to design to make it easy to adopt the behaviors. Those early adopters are critical for achieving mainstream adoption based on their direct influence on peers and the indirect influence of visibility of their behaviors, which together can make it normative to adopt the behaviors. Bottom-up behavior change will eventually stall if there is not stakeholder support to shift incentives to make it desirable to adopt the behaviors. And even incentives may not be sufficient to unseat behaviors that are sustained by structural factors: Policy changes can then adapt the structure and make it required to adopt the behaviors.

The model's five levels of intervention are highly interdependent; each is necessary and none is sufficient to achieve culture and behavior changes. For example, a policy intervention that does not have quality infrastructure or normative support will be perceived as an unwelcome bureaucratic burden and is likely to fail to meet its goals.

Evidence of Change

Behaviors that may directly or indirectly improve replicability, or the ability to assess replicability, include increasing sample size, preregistering studies, improving rigor and transparency, sharing materials and primary data, conducting replications, and enhancing error detection and correction. A variety of interventions and solutions have emerged in the last decade, including tools supporting preregistration and sharing, such as the Open Science Framework (OSF; see Soderberg 2018) and AsPredicted, and facilitating error detection and correction, such as statcheck (Epskamp & Nuijten 2018) and granularity-related inconsistency of means (GRIM; see Brown & Heathers 2017); grassroots communities promoting new norms, such as the Society for the Improvement of Psychological Science, open science communities (Armeni et al. 2020), and national reproducibility networks (Munafò et al. 2020); large-scale collaborations to increase sample size and replication efforts, such as the Psychological Science Accelerator (Moshontz et al. 2018) and ManyBabies (Byers-Heinlein et al. 2020); badges for open practices and other ways to shift norms by increasing the visibility of desirable behaviors (Kidwell et al. 2016); new incentives for publishing not only positive, novel, and tidy results (e.g., with Registered Reports) (Chambers 2019, Scheel et al. 2020); and policy changes by publishers, funders, and institutions to encourage or require more rigor, transparency, and sharing [e.g., the Transparency and Openness Promotion (TOP) Guidelines] (Nosek et al. 2015).

Most available survey data suggest that psychologists and other social-behavioral researchers acknowledge engaging in questionable research practices that could interfere with replicability (John et al. 2012). **Supplemental Table 5** summarizes 14 surveys of questionable research practices (QRPs) ($N = 7{,}887$), showing, for example, that 9% to 43% of researchers acknowledged failing to report all study outcomes, and 25% to 62% of researchers acknowledged selectively reporting studies that "worked." We cannot surmise from these data whether QRPs are changing over time, both because of variation in sampling strategies and because most surveys asked if researchers had *ever* engaged in the behaviors. However, the median occurrence across surveys suggests that many of these behaviors are relatively common.

Supplemental Material >

Seven surveys of researchers ($N = 4{,}737$; see **Supplemental Table 3**) and four audit studies (of a total of 1,100 articles; see **Supplemental Table 4**) assessed behaviors like sharing data and materials or preregistering studies among psychological scientists. The surveys observed that between 27% and 60% of researchers reporting having shared data, and between 27% and 57%

reporting having preregistered a study. The audit studies observed high variation in data sharing (found in 0–65% of articles), likely due to the sampling strategy used, and only one study assessed preregistration and observed a rate of 3%.

The audit studies suggest that self-reported behaviors have not yet translated into significant changes in the published literature; for example, only 2% of a random sample of psychology studies published between 2014 and 2017 had shared data, and only 3% had been preregistered (Hardwicke et al. 2020b). The discrepancy between audits and self-reports is likely a function of multiple factors, including when the surveys and audits were conducted, the possible over-reporting of certain behaviors in the surveys, the time lag between the moment a behavior is performed and the moment it is reflected in a published article, and the fact that surveys tend to ask about having a certain behavior once, whereas individual researchers conduct multiple studies. Continuing issues with publication bias also imply that not all newly conducted studies are published.

Christensen and colleagues (2019) asked psychologists to retrospectively report when they first preregistered a study or posted data or code online, and they observed that about 20% of participants had shared data or code and about 8% had preregistered studies in 2011, with those numbers rising to 51% and 44%, respectively, by 2017. Supporting that self-reported evidence, the usage of services like OSF and AsPredicted for preregistration and the sharing of data and materials has grown exponentially (**Figure 4**). Both services are available to anyone, but a substantial portion of their users are from psychology and allied fields. A 2019 analysis of all faculty from 69 psychology departments ($N = 1{,}987$) indicated that 35% of the participants had OSF accounts, with the heaviest representation in social (57%), quantitative (48%), and cognitive (42%) psychology and the lightest representation in clinical (19%) and education and health (17%) psychology (Nosek 2019b).

Supplemental Material >

Contrasting the evidence of behavior change, in the **Supplemental Text** we summarize five investigations of sample size over time and do not observe compelling evidence of change from 1977

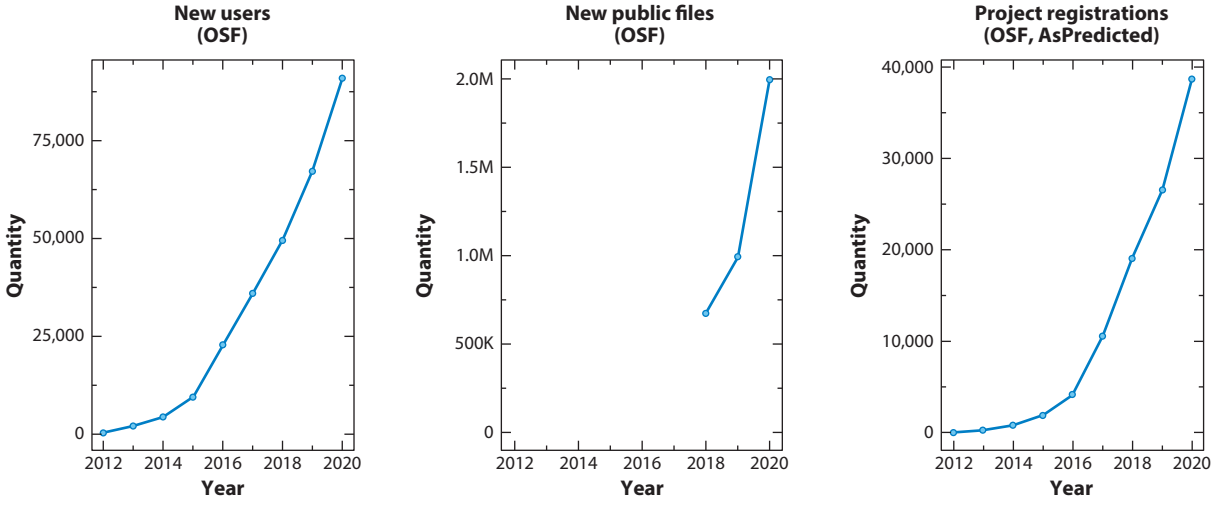

Figure 4

Yearly counts of users, sharing of files (research data, materials, code), and registration of studies on OSF and AsPredicted, two popular services for psychologists and allied disciplines. Data for new public files (sharing) prior to 2018 are not available. Abbreviation: OSF, Open Science Framework.

to 2017; studies of statistical reporting errors over time also suggest relative stability from 1985 to 2013 (Bakker & Wicherts 2011, Nuijten et al. 2016). Investigations of sample size and statistical errors require updated investigations for the latter half of the decade. Also, in the **Supplemental Text** we report evidence that retractions in psychology are still rare but increasing. The cause is not clear, but a plausible explanation is greater attention to and effort toward the detection and correction of misconduct, faulty research, and honest errors (Marcus & Oransky 2018).

Supplemental Material >

Evidence of individual behavior change is complemented by the adoption of norms, incentives, and policy interventions by journals and other stakeholder groups. Idiosyncratic actions by stakeholders have given prominence to replicability; for example, the Netherlands Organization for Scientific Research, the National Science Foundation, and the Defense Advanced Research Projects Agency (DARPA) have issued calls for replication research proposals (Baker 2016, Cook 2016, Wiktop 2020). The German Research Foundation launched a meta-scientific program to analyze and optimize replicability in the behavioral, social, and cognitive sciences (Gollwitzer 2020), and individual institutions and departments articulated principles for improving rigor and replicability or expressed interest in incorporating such factors into hiring and promotion. We conducted two systematic inquiries to assess the current status of journal policies and psychology department hiring practices.

The TOP Guidelines (**https://cos.io/top/**) are a set of 10 policy standards related to transparency and reproducibility, each with three levels of increasing stringency (Nosek et al. 2015) (see **Supplemental Table 7**). We assessed the adoption of TOP-compliant policies in a random sample of psychology journals ($N = 40$) and in the five journals with the highest impact factor from each of 10 psychology subfields ($N = 50$). Methodological details are available in the **Supplemental Text**. As illustrated in **Figure 5**, for each of the 10 standards, the substantial majority of journals had not adopted TOP-compliant policies (i.e., policy level 0 was found in a range of 56–90% journals, with a median of 83%). For 8 of the 10 standards, high-impact journals were more likely than random journals to have adopted a policy at any level, though the overall frequency of policy adoption was comparable between the two types of journals (17% and 15%, respectively). Combining samples, TOP-compliant policies were most common for citing data sources (36%) and using reporting guidelines (36%), and they were least common for preregistration of studies (12%) and analysis plans (13%). These findings suggest a modest adoption of replicability-related policies among psychology journals. Notably, psychology's largest publisher, the American Psychological Association, has indicated its intention to move all of its core journals to at least policy level 1 across eight standards by the end of 2021 (Cent. Open Sci. 2020).

We also examined whether research institutions are explicitly communicating expectations of replicability and transparency in their job advertisements. We analyzed all academic job offers in psychology from the German Academics platform (**https://www.academics.de/**) from February 2017 to December 2020 ($N = 1{,}626$). Overall, 2.2% ($N = 36$) of job offers mentioned replicability and transparency as desired or essential job criteria. Most of these mentions ($N = 24$) concerned professorship positions, whereas the remainder ($N = 12$) concerned other scientific personnel. Of 376 advertising institutions, 20 mentioned replicability and transparency at least once. These numbers are small, but there are hints of an increasing trend (in 2017 and 2018, the percentage of advertising institutions mentioning replicability was 1.0%; in 2019, 2.0%; in 2020, 3.8%).

There is both substantial evidence of new behaviors that may increase the rigor and replicability of psychological findings and substantial evidence that more work is needed to address the structural, cultural, social, and individual barriers to change. So far, the driver of change has been the grassroots efforts by individuals and groups to improve research practices. Journals are

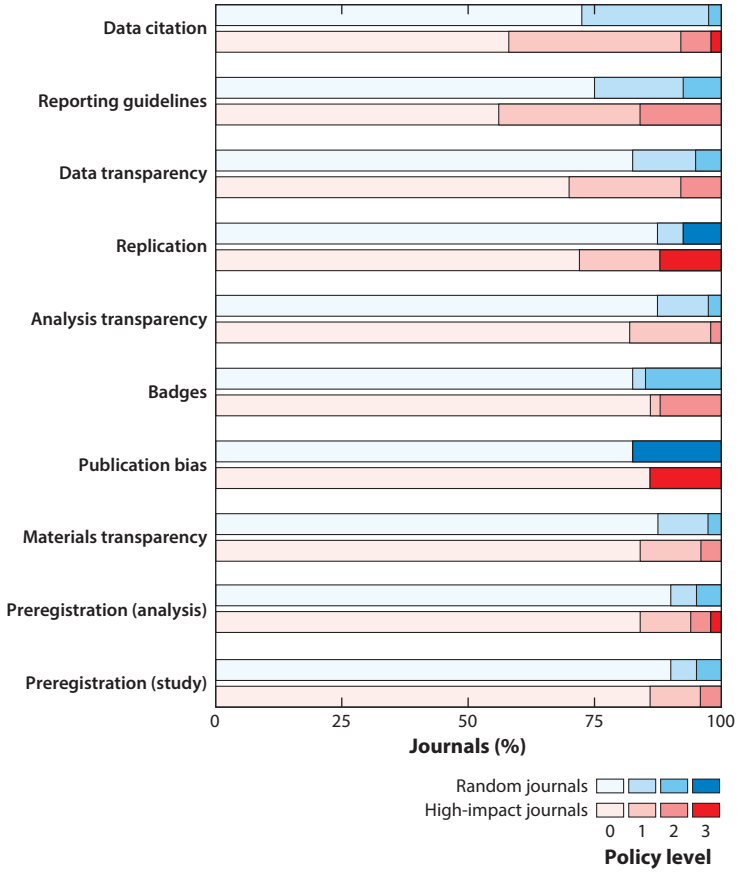

Figure 5

Adoption of Transparency and Openness Promotion (TOP) policies by randomly selected ($N = 40$) (*blue*) and high-impact ($N = 50$) (*red*) psychology journals. Policies are ordered by the proportion of journals adopting the policy at any level.

leading change among stakeholder groups, with department and institutional practices for hiring and promotion showing the least evidence of change so far.

WHAT'S NEXT? A METASCIENCE RESEARCH AND CULTURE CHANGE AGENDA FOR ACCELERATING PSYCHOLOGICAL SCIENCE

Like any good program of research, the productive decade of research on replicability has brought important questions to the fore that will be fodder for the next decade of metascience research (Hardwicke et al. 2020a, Zwaan et al. 2018). First, what is the optimal replicability rate at different stages of research maturity? How do we maximize progress and minimize waste (Lewandowsky & Oberauer 2020, Shiffrin et al. 2018)? And what role do behaviors that promote replicability play in that optimization process? There is not yet good evidence about these questions.

Second, what is the role of replicability in building cumulative science? Replicability is one of a variety of topics that are relevant for the credibility of research findings and the translation of knowledge into application. Other issues include measurement, causal inference, theory,

generalizability, and applicability. These topics are interdependent but not redundant. Replicability does not guarantee validity of measurement or causal inference, nor that the knowledge is applicable. Theorists vary in their weighting of which areas need to be improved in order to advance knowledge (Devezer et al. 2019, Feest 2019, Frank et al. 2017, Leonelli 2018). And, at present, there is little empirical evidence to advance these debates.

Third, are interventions to improve replicability effective? The earlier sections examining what replicates and opportunities for improving replicability provided a reasonable conceptual basis for believing that interventions such as increasing sample size, improving formalization of generating hypotheses, and preregistering studies and analysis plans will improve replicability. However, there is too little empirical data to verify whether this is the case. An immediate research priority is to evaluate the variety of interventions that are gaining traction in psychological science.

Finally, what is working, what is not, and what is still needed in the culture change movement? Interventions to improve inclusivity, reward systems, error detection, and team science have gained momentum, but are they actually changing the research culture? And are they improving the research culture, or are they having unintended negative consequences that outweigh the intended benefits? A healthy metascience and culture change movement will be constantly evaluating their progress and impact to adapt and change course as demanded by the evidence.

Replication can prompt challenge and uncertainty, even acrimony. However, when replication is incorporated as an ordinary part of skeptical inquiry, the occasional acrimony can be eclipsed by feelings of excitement, empowerment, and enlightenment. Replicability and credibility challenges have been recognized for decades with little to no evidence of change. Now things are changing. There is much more to do, but the hardest part is getting started. That part is done.

DISCLOSURE STATEMENT

B.A.N. and M.K.S. are employees of the nonprofit Center for Open Science that has a mission to increase openness, integrity, and reproducibility of research. K.S.C. is the unpaid executive officer of the nonprofit Society for the Improvement of Psychological Science. J.M.R. is a member of the Society for the Improvement of Psychological Science and a board member of two journals dedicated to open science practices, *Meta-Psychology* and *Personality Science*. M.B.N. is a member of the Society for the Improvement of Psychological Science, a board member of four journals dedicated to open science practices (*Collabra: Psychology*, *Scientific Reports*, *Meta-Psychology*, and *Personality Science*), and a member of the Science Committee at Tilburg University, aimed at facilitating and increasing data sharing.

A.M.S. is a member of the Society for the Improvement of Psychological Science and of the Open Science Committee of the German Psychological Society (DGPs) and serves as an advisory board member for the UK Reproducibility Network and for the Collaborative Assessment for Trustworthy Science project (repliCATS) at the University of Melbourne. L.D.S. is a founding member of the Society for the Improvement of Psychological Science. F.D.S. is managing director of the Ludwig Maximilian University Open Science Center. The other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

AUTHOR CONTRIBUTIONS

B.A.N. drafted the outline and sections of the manuscript, collaborated with section leads on conceptualizing and drafting their components, coded replication study outcomes for **Figure 1**, and coded journal policies for **Figure 5**. T.E.H. drafted the section titled Evidence of Change and drafted the section titled What Happens After Replication? in the **Supplemental Text**,

Supplemental Material >

collected and analyzed data for **Figure 5** and **Supplemental Figure 1**, and made suggestions and revisions to the manuscript prior to submission. H.M. contributed content to the section titled Evidence of Change; drafted the section titled Are Behaviors Changing? in the **Supplemental Text**; compiled, curated, and synthesized data for **Supplemental Tables 4–6**; and contributed to curating data for **Figure 1**. A.A. drafted small sections of the manuscript and compiled, curated, and analyzed data for **Figure 1**. A.D. drafted small sections of the manuscript and compiled and analyzed data for **Figure 2** and **Supplemental Figure 2**. J.H. drafted parts of the section titled Evidence of Change and the section titled Retractions in the **Supplemental Text**, analyzed data for these sections, and generated **Figure 4**. F.R. drafted small sections of the manuscript and parts of the section titled What Happens After Replication? in the **Supplemental Text** and made revisions to the manuscript. L.D.S. drafted parts of the section titled Evidence of Change; contributed to the literature review and to coding appearing in **Supplemental Tables 4 and 5** and in **Figure 4**; and made suggestions and revisions to the manuscript prior to submission. F.D.S. drafted small sections of the manuscript and collected and analyzed data for the section titled Changes in Job Advertisements in the **Supplemental Text**. K.S.C., F.F., M.K.S., M.B.N., J.M.R., A.M.S., and S.V. drafted small sections of the manuscript and made revisions to the manuscript.

ACKNOWLEDGMENTS

This work was supported by grants to B.A.N. from Arnold Ventures, the John Templeton Foundation, Templeton World Charity Foundation, and Templeton Religion Trust. T.E.H. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 841188. F.F. is funded by an Australian Research Council Future Fellowship (FT150100297). A.M.S. is currently funded through a grant by the Dutch Research Council (NWO) on "Increasing the Reliability and Efficiency of Psychological Science." L.D.S. has received funding from the National Institutes of Health, the Agency for Healthcare Research and Quality, and the Patient-Centered Outcomes Research Institute. We thank Adam Gill for assistance creating **Figures 1** and **2** and **Supplemental Figure 2**. Data, materials, and code are available at **https://osf.io/7np92/**.

LITERATURE CITED

Alogna VK, Attaya MK, Aucoin P, Bahník Š, Birch S, et al. 2014. Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.* 9(5):556–78

Altmejd A, Dreber A, Forsell E, Huber J, Imai T, et al. 2019. Predicting the replicability of social science lab experiments. *PLOS ONE* 14(12):e0225826

Anderson CJ, Bahník Š, Barnett-Cowan M, Bosco FA, Chandler J, et al. 2016. Response to Comment on "Estimating the reproducibility of psychological science." *Science* 351(6277):1037

Anderson MS, Martinson BC, De Vries R. 2007. Normative dissonance in science: results from a national survey of U.S. scientists. *J. Empir. Res. Hum. Res. Ethics* 2(4):3–14

Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. 2018. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. *Am. Psychol.* 73(1):3–25. Corrigendum. 2018. *Am. Psychol.* 73(7):947

Armeni K, Brinkman L, Carlsson R, Eerland A, Fijten R, et al. 2020. Towards wide-scale adoption of open science practices: the role of open science communities. MetaArXiv, Oct. 6. **https://doi.org/10.31222/osf.io/7gct9**

Artner R, Verliefde T, Steegen S, Gomes S, Traets F, et al. 2020. The reproducibility of statistical results in psychological research: an investigation using unpublished raw data. *Psychol. Methods.* In press. **https://doi.org/10.1037/met0000365**

Baker M. 2016. Dutch agency launches first grants programme dedicated to replication. *Nat. News.* **https://doi.org/10.1038/nature.2016.20287**

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

Baribault B, Donkin C, Little DR, Trueblood JS, Oravecz Z, et al. 2018. Metastudies for robust tests of theory. *PNAS* 115(11):2607–12

Baron J, Hershey JC. 1988. Outcome bias in decision evaluation. *J. Pers. Soc. Psychol.* 54(4):569–79

Baumeister RF. 2016. Charting the future of social psychology on stormy seas: winners, losers, and recommendations. *J. Exp. Soc. Psychol.* 66:153–58

Baumeister RF, Vohs KD. 2016. Misguided effort with elusive implications. *Perspect. Psychol. Sci.* 11(4):574–75

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, et al. 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2(1):6–10

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582(7810):84–88

Bouwmeester S, Verkoeijen PPJL, Aczel B, Barbosa F, Bègue L, et al. 2017. Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspect. Psychol. Sci.* 12(3):527–42

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

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

Byers-Heinlein K, Bergmann C, Davies C, Frank M, Hamlin JK, et al. 2020. Building a collaborative psychological science: lessons learned from ManyBabies 1. *Can. Psychol. Psychol. Can.* 61(4):349–63

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

Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. 2019. Correcting for bias in psychology: a comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* 2(2):115–44

Cent. Open Sci. 2020. APA joins as new signatory to TOP guidelines. *Center for Open Science*, Nov. 10. **https://www.cos.io/about/news/apa-joins-as-new-signatory-to-top-guidelines**

Cesario J. 2014. Priming, replication, and the hardest science. *Perspect. Psychol. Sci.* 9(1):40–48

Chambers C. 2019. What's next for Registered Reports? *Nature* 573(7773):187–89

Cheung I, Campbell L, LeBel EP, Ackerman RA, Aykutoğlu B, et al. 2016. Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspect. Psychol. Sci.* 11(5):750–64

Christensen G, Wang Z, Paluck EL, Swanson N, Birke DJ, Miguel E, Littman R. 2019. Open science practices are on the rise: the State of Social Science (3S) Survey. MetaArXiv, Oct. 18. **https://doi.org/10.31222/osf.io/5rksu**

Christensen-Szalanski JJ, Willham CF. 1991. The hindsight bias: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 48(1):147–68

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65(3):145–53

Cohen J. 1973. Statistical power analysis and research results. *Am. Educ. Res. J.* 10(3):225–29

Cohen J. 1990. Things I have learned (so far). *Am. Psychol.* 45:1304–12

Cohen J. 1992. A power primer. *Psychol. Bull.* 112(1):155–59

Cohen J. 1994. The earth is round ($p < .05$). *Am. Psychol.* 49(12):997–1003

Colling LJ, Szücs D, De Marco D, Cipora K, Ulrich R, et al. 2020. Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003). *Adv. Methods Pract. Psychol. Sci.* 3(2):143–62

Cook FL. 2016. Dear Colleague Letter: robust and reliable research in the social, behavioral, and economic sciences. *National Science Foundation*, Sept. 20. **https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp**

Crandall CS, Sherman JW. 2016. On the scientific superiority of conceptual replications for scientific progress. *J. Exp. Soc. Psychol.* 66:93–99

Crisp RJ, Miles E, Husnu S. 2014. Support for the replicability of imagined contact effects. *Soc. Psychol.* 45(4):303–4

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52(4):281–302

Dang J, Barker P, Baumert A, Bentvelzen M, Berkman E, et al. 2021. A multilab replication of the ego depletion effect. *Soc. Psychol. Pers. Sci.* 12(1):14–24

Devezer B, Nardin LG, Baumgaertner B, Buzbas EO. 2019. Scientific discovery in a model-centric framework: reproducibility, innovation, and epistemic diversity. *PLOS ONE* 14(5):e0216125

Dijksterhuis A. 2018. Reflection on the professor-priming replication report. *Perspect. Psychol. Sci.* 13(2):295–96

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

Duhem PMM. 1954. *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton Univ. Press

Ebersole CR, Alaei R, Atherton OE, Bernstein MJ, Brown M, et al. 2017. Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science. *J. Exp. Soc. Psychol.* 69:184–86

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, et al. 2016a. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* 67:68–82

Ebersole CR, Axt JR, Nosek BA. 2016b. Scientists' reputations are based on getting it right, not being right. *PLOS Biol.* 14(5):e1002460

Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, et al. 2020. Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Adv. Methods Pract. Psychol. Sci.* 3(3):309–31

Eerland A, Sherrill AM, Magliano JP, Zwaan RA, Arnal JD, et al. 2016. Registered Replication Report: Hart & Albarracín (2011). *Perspect. Psychol. Sci.* 11(1):158–71

Ellemers N, Fiske ST, Abele AE, Koch A, Yzerbyt V. 2020. Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *PNAS* 117(14):7561–67

Epskamp S, Nuijten MB. 2018. Statcheck: extract statistics from articles and recompute p values. *Statistical Software.* **https://CRAN.R-project.org/package=statcheck**

Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. 2021. Challenges for assessing reproducibility and replicability in preclinical cancer biology. *eLife.* In press

Etz A, Vandekerckhove J. 2016. A Bayesian perspective on the reproducibility project: psychology. *PLOS ONE* 11(2):e0149794

Fanelli D. 2010. "Positive" results increase down the hierarchy of the sciences. *PLOS ONE* 5(4):e10068

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3):891–904

Feest U. 2019. Why replication is overrated. *Philos. Sci.* 86(5):895–905

Ferguson MJ, Carter TJ, Hassin RR. 2014. Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Soc. Psychol.* 45(4):301–2

Fetterman AK, Sassenberg K. 2015. The reputational consequences of failed replications and wrongness admission among scientists. *PLOS ONE* 10(12):e0143723

Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, et al. 2019. Predicting replication outcomes in the Many Labs 2 study. *J. Econ. Psychol.* 75:102117

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: evidence from a study registry. *Soc. Psychol. Pers. Sci.* 7(1):8–12

Frank MC, Bergelson E, Bergmann C, Cristia A, Floccia C, et al. 2017. A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22(4):421–35

Funder DC, Ozer DJ. 2019. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* 2(2):156–68

Gelman A, Carlin J. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* 9(6):641–51

Gelman A, Loken E. 2013. *The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Work. Pap., Columbia Univ., New York

Gergen KJ. 1973. Social psychology as history. *J. Pers. Soc. Psychol.* 26(2):309–20
Gervais WM, Jewell JA, Najle MB, Ng BKL. 2015. A powerful nudge? Presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs. *Soc. Psychol. Pers. Sci.* 6(7):847–54
Ghelfi E, Christopherson CD, Urry HL, Lenne RL, Legate N, et al. 2020. Reexamining the effect of gustatory disgust on moral judgment: a multilab direct replication of Eskine, Kacinik, and Prinz (2011). *Adv. Methods Pract. Psychol. Sci.* 3(1):3–23
Gilbert DT, King G, Pettigrew S, Wilson TD. 2016. Comment on "Estimating the reproducibility of psychological science." *Science* 351(6277):1037
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
Giner-Sorolla R. 2019. From crisis of evidence to a "crisis" of relevance? Incentive-based answers for social psychology's perennial relevance worries. *Eur. Rev. Soc. Psychol.* 30(1):1–38
Gollwitzer M. 2020. DFG Priority Program SPP 2317 Proposal: A meta-scientific program to analyze and optimize replicability in the behavioral, social, and cognitive sciences (META-REP). PsychArchives, May 29. **http://dx.doi.org/10.23668/psycharchives.3010**
Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, et al. 2020. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R. Soc. Open Sci.* 7(7):200566
Götz M, O'Boyle EH, Gonzalez-Mulé E, Banks GC, Bollmann SS. 2020. The "Goldilocks Zone": (Too) many confidence intervals in tests of mediation just exclude zero. *Psychol. Bull.* 147(1):95–114
Greenwald AG. 1975. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* 82(1):1–20
Hagger MS, Chatzisarantis NLD, Alberts H, Anggono CO, Batailler C, et al. 2016. A multilab preregistered replication of the ego-depletion effect. *Perspect. Psychol. Sci.* 11(4):546–73
Hanea AM, McBride MF, Burgman MA, Wintle BC, Fidler F, et al. 2017. I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *Int. J. Forecast.* 33(1):267–79
Hardwicke TE, Bohn M, MacDonald KE, Hembacher E, Nuijten MB, et al. 2021. Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: an observational study. *R. Soc. Open Sci.* 8(1):201494
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, Serghiou S, Janiaud P, Danchev V, Crüwell S, et al. 2020a. Calibrating the scientific ecosystem through meta-research. *Annu. Rev. Stat. Appl.* 7:11–37
Hardwicke TE, Thibault RT, Kosie JE, Wallach JD, Kidwell M, Ioannidis J. 2020b. Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). MetaArXiv, Jan. 2. **https://doi.org/10.31222/osf.io/9sz2y**
Hedges LV, Schauer JM. 2019. Statistical analyses for studying replication: meta-analytic perspectives. *Psychol. Methods* 24(5):557–70
Hoogeveen S, Sarafoglou A, Wagenmakers E-J. 2020. Laypeople can predict which social-science studies will be replicated successfully. *Adv. Methods Pract. Psychol. Sci.* 3(3):267–85
Hughes BM. 2018. *Psychology in Crisis*. London: Palgrave Macmillan
Inbar Y. 2016. Association between contextual dependence and replicability in psychology may be spurious. *PNAS* 113(34):E4933–34
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. 2014. How to make more published research true. *PLOS Med.* 11(10):e1001747
Ioannidis JPA, Trikalinos TA. 2005. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol.* 58(6):543–49
Isager PM, van Aert RCM, Bahník Š, Brandt M, DeSoto KA, et al. 2020. Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection. MetaArXiv, Sept. 2. **https://doi.org/10.31222/osf.io/2gurz**

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

Kahneman D. 2003. Experiences of collaborative research. *Am. Psychol.* 58(9):723–30

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

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, et al. 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLOS Biol.* 14(5):e1002456

Klein RA, Cook CL, Ebersole CR, Vitiello C, Nosek BA, et al. 2019. Many Labs 4: failure to replicate mortality salience effect with and without original author involvement. PsyArXiv, Dec. 11. **https://doi.org/10/ghwq2w**

Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, et al. 2014. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* 45(3):142–52

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

Kunda Z. 1990. The case for motivated reasoning. *Psychol. Bull.* 108(3):480–98

Lakens D. 2019. The value of preregistration for psychological science: a conceptual analysis. PsyArXiv, Nov. 18. **https://doi.org/10.31234/osf.io/jbh4w**

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MA, et al. 2018. Justify your alpha. *Nat. Hum. Behav.* 2(3):168–71

Landy JF, Jia ML, Ding IL, Viganola D, Tierney W, et al. 2020. Crowdsourcing hypothesis tests: making transparent how design choices shape research results. *Psychol. Bull.* 146(5):451–79

Leary MR, Diebels KJ, Davisson EK, Jongman-Sereno KP, Isherwood JC, et al. 2017. Cognitive and interpersonal features of intellectual humility. *Pers. Soc. Psychol. Bull.* 43(6):793–813

LeBel EP, McCarthy RJ, Earp BD, Elson M, Vanpaemel W. 2018. A unified framework to quantify the credibility of scientific findings. *Adv. Methods Pract. Psychol. Sci.* 1(3):389–402

Leighton DC, Legate N, LePine S, Anderson SF, Grahe J. 2018. Self-esteem, self-disclosure, self-expression, and connection on Facebook: a collaborative replication meta-analysis. *Psi Chi J. Psychol. Res.* 23(2):98–109

Leising D, Thielmann I, Glöckner A, Gärtner A, Schönbrodt F. 2020. Ten steps toward a better personality science—how quality may be rewarded more in research evaluation. PsyArXiv, May 31. **https://doi.org/10.31234/osf.io/6btc3**

Leonelli S. 2018. Rethinking reproducibility as a criterion for research quality. In *Research in the History of Economic Thought and Methodology*, Vol. 36, ed. L Fiorito, S Scheall, CE Suprinyak, pp. 129–46. Bingley, UK: Emerald

Lewandowsky S, Oberauer K. 2020. Low replicability can support robust and efficient science. *Nat. Commun.* 11(1):358

Maassen E, van Assen MALM, Nuijten MB, Olsson-Collentine A, Wicherts JM. 2020. Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE* 15(5):e0233107

Machery E. 2020. What is a replication? *Philos. Sci.* 87(4):545–67

ManyBabies Consort. 2020. Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3(1):24–52

Marcus A, Oransky I. 2018. Meet the "data thugs" out to expose shoddy and questionable research. *Science*, Feb. 18. **https://www.sciencemag.org/news/2018/02/meet-data-thugs-out-expose-shoddy-and-questionable-research**

Marcus A, Oransky I. 2020. Tech firms hire "Red Teams." Scientists should, too. *WIRED*, July 16. **https://www.wired.com/story/tech-firms-hire-red-teams-scientists-should-too/**

Mathur MB, VanderWeele TJ. 2020. New statistical metrics for multisite replication projects. *J. R. Stat. Soc. A* 183(3):1145–66

Maxwell SE. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods* 9(2):147–63

Maxwell SE, Lau MY, Howard GS. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70(6):487–98

Mayo DG. 2018. *Statistical Inference as Severe Testing*. Cambridge, UK: Cambridge Univ. Press

McCarthy R, Gervais W, Aczel B, Al-Kire R, Baraldo S, et al. 2021. A multi-site collaborative study of the hostile priming effect. *Collabra Psychol.* 7(1):18738

McCarthy RJ, Hartnett JL, Heider JD, Scherer CR, Wood SE, et al. 2018. An investigation of abstract construal on impression formation: a multi-lab replication of McCarthy and Skowronski (2011). *Int. Rev. Soc. Psychol.* 31(1):15

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46(4):806–34

Meyer MN, Chabris C. 2014. Why psychologists' food fight matters. *Slate Magazine*, July 31. **https://slate.com/technology/2014/07/replication-controversy-in-psychology-bullying-file-drawer-effect-blog-posts-repligate.html**

Mischel W. 2008. The toothbrush problem. *APS Observer*, Dec. 1. **https://www.psychologicalscience.org/observer/the-toothbrush-problem**

Moran T, Hughes S, Hussey I, Vadillo MA, Olson MA, et al. 2020. Incidental attitude formation via the surveillance task: a Registered Replication Report of Olson and Fazio (2001). PsyArXiv, April 17. **https://doi.org/10/ghwq2z**

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

Munafò MR, Chambers CD, Collins AM, Fortunato L, Macleod MR. 2020. Research culture and reproducibility. *Trends Cogn. Sci.* 24(2):91–93

Muthukrishna M, Henrich J. 2019. A problem in theory. *Nat. Hum. Behav.* 3(3):221–29

Natl. Acad. Sci. Eng. Med. 2019. *Reproducibility and Replicability in Science*. Washington, DC: Natl. Acad. Press

Nelson LD, Simmons J, Simonsohn U. 2018. Psychology's renaissance. *Annu. Rev. Psychol.* 69:511–34

Nickerson RS. 1998. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2(2):175–220

Nosek B. 2019a. Strategy for culture change. *Center for Open Science*, June 11. **https://www.cos.io/blog/strategy-for-culture-change**

Nosek B. 2019b. The rise of open science in psychology, a preliminary report. *Center for Open Science*, June 3. **https://www.cos.io/blog/rise-open-science-psychology-preliminary-report**

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, Beck ED, Campbell L, Flake JK, Hardwicke TE, et al. 2019. Preregistration is hard, and worthwhile. *Trends Cogn. Sci.* 23(10):815–18

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *PNAS* 115(11):2600–6

Nosek BA, Errington TM. 2020a. What is replication? *PLOS Biol.* 18(3):e3000691

Nosek BA, Errington TM. 2020b. The best time to argue about what a replication means? Before you do it. *Nature* 583(7817):518–20

Nosek BA, Gilbert EA. 2017. Mischaracterizing replication studies leads to erroneous conclusions. PsyArXiv, April 18. **https://doi.org/10.31234/osf.io/nt4d3**

Nosek BA, Spies JR, Motyl M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. On Psychol. Sci.* 7(6):615–31

Nuijten MB, Bakker M, Maassen E, Wicherts JM. 2018. Verify original results through reanalysis before replicating. *Behav. Brain Sci.* 41:e143

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

Nuijten MB, van Assen MA, Veldkamp CL, Wicherts JM. 2015. The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Rev. Gen. Psychol.* 19(2):172–82

O'Donnell M, Nelson LD, Ackermann E, Aczel B, Akhtar A, et al. 2018. Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspect. Psychol. Sci.* 13(2):268–94

Olsson-Collentine A, Wicherts JM, van Assen MALM. 2020. Heterogeneity in direct replications in psychology and its association with effect size. *Psychol. Bull.* 146(10):922–40

Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
Patil P, Peng RD, Leek JT. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11(4):539–44
Pawel S, Held L. 2020. Probabilistic forecasting of replication studies. *PLOS ONE* 15(4):e0231416
Perugini M, Gallucci M, Costantini G. 2014. Safeguard power as a protection against imprecise power estimates. *Perspect. Psychol. Sci.* 9(3):319–32
Protzko J, Krosnick J, Nelson LD, Nosek BA, Axt J, et al. 2020. High replicability of newly-discovered social-behavioral findings is achievable. PsyArXiv, Sept. 10. **https://doi.org/10.31234/osf.io/n2a9x**
Rogers EM. 2003. *Diffusion of Innovations*. New York: Free Press. 5th ed.
Romero F. 2017. Novelty versus replicability: virtues and vices in the reward system of science. *Philos. Sci.* 84(5):1031–43
Rosenthal R. 1979. The file drawer problem and tolerance for null results. *Psychol. Bull.* 86(3):638–41
Rothstein HR, Sutton AJ, Borenstein M. 2005. Publication bias in meta-analysis. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, ed. HR Rothstein, AJ Sutton, M Borenstein, pp. 1–7. Chichester, UK: Wiley & Sons
Rouder JN. 2016. The what, why, and how of born-open data. *Behav. Res. Methods* 48(3):1062–69
Scheel AM, Schijen M, Lakens D. 2020. An excess of positive results: comparing the standard psychology literature with Registered Reports. PsyArXiv, Febr. 5. **https://doi.org/10.31234/osf.io/p6e9c**
Schimmack U. 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17(4):551–66
Schmidt S. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13(2):90–100
Schnall S. 2014. Commentary and rejoinder on Johnson, Cheung, and Donnellan (2014a). Clean data: Statistical artifacts wash out replication efforts. *Soc. Psychol.* 45(4): 315–17
Schwarz N, Strack F. 2014. Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Soc. Psychol.* 45(4):305–6
Schweinsberg M, Madan N, Vianello M, Sommer SA, Jordan J, et al. 2016. The pipeline project: pre-publication independent replications of a single laboratory's research pipeline. *J. Exp. Soc. Psychol.* 66:55–67
Sedlmeier P, Gigerenzer G. 1992. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105(2):309–16
Shadish WR, Cook TD, Campbell DT, eds. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin
Shiffrin RM, Börner K, Stigler SM. 2018. Scientific progress despite irreproducibility: a seeming paradox. *PNAS* 115(11):2632–39
Shih M, Pittinsky TL. 2014. Reflections on positive stereotypes research and on replications. *Soc. Psychol.* 45(4):335–38
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
Simons DJ. 2014. The value of direct replication. *Perspect. Psychol. Sci.* 9(1):76–80
Simons DJ, Shoda Y, Lindsay DS. 2017. Constraints on generality (COG): a proposed addition to all empirical papers. *Perspect. Psychol. Sci.* 12(6):1123–28
Simonsohn U. 2015. Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* 26(5):559–69
Simonsohn U, Simmons JP, Nelson LD. 2020. Specification curve analysis. *Nat. Hum. Behav.* 4:1208–14
Smaldino PE, McElreath R. 2016. The natural selection of bad science. *R. Soc. Open Sci.* 3(9):160384
Smith PL, Little DR. 2018. Small is beautiful: in defense of the small-N design. *Psychon. Bull. Rev.* 25(6):2083–101
Soderberg CK. 2018. Using OSF to share data: a step-by-step guide. *Adv. Methods Pract. Psychol. Sci.* 1(1):115–20

Soderberg CK, Errington T, Schiavone SR, Bottesini JG, Thorn FS, et al. 2021. Initial evidence of research quality of Registered Reports compared with the standard publishing model. *Nat. Hum. Behav.* 5(8):990–97

Soto CJ. 2019. How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychol. Sci.* 30(5):711–27

Spellman BA. 2015. A short (personal) future history of revolution 2.0. *Perspect. Psychol. Sci.* 10(6):886–99

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

Sterling TD, Rosenbaum WL, Weinkam JJ. 1995. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Am. Stat.* 49:108–12

Stroebe W, Strack F. 2014. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9(1):59–71

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

Tiokhin L, Derex M. 2019. Competition for novelty reduces information sampling in a research game—a registered report. *R. Soc. Open Sci.* 6(5):180934

Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. 2016. Contextual sensitivity in scientific reproducibility. *PNAS* 113(23):6454–59

Vazire S. 2018. Implications of the credibility revolution for productivity, creativity, and progress. *Perspect. Psychol. Sci.* 13(4):411–17

Vazire S, Schiavone SR, Bottesini JG. 2020. Credibility beyond replicability: improving the four validities in psychological science. PsyArXiv, Oct. 7. **https://doi.org/10.31234/osf.io/bu4d3**

Verhagen J, Wagenmakers E-J. 2014. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* 143(4):1457–75

Verschuere B, Meijer EH, Jim A, Hoogesteyn K, Orthey R, et al. 2018. Registered Replication Report on Mazar, Amir, and Ariely (2008). *Adv. Methods Pract. Psychol. Sci.* 1(3):299–317

Vosgerau J, Simonsohn U, Nelson LD, Simmons JP. 2019. 99% impossible: a valid, or falsifiable, internal meta-analysis. *J. Exp. Psychol. Gen.* 148(9):1628–39

Wagenmakers E-J, Beek T, Dijkhoff L, Gronau QF, Acosta A, et al. 2016. Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspect. Psychol. Sci.* 11(6):917–28

Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas HL. 2011. Why psychologists must change the way they analyze their data. The case of psi: comment on Bem (2011). *J. Pers. Soc. Psychol.* 100(3):426–32

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

Wagge J, Baciu C, Banas K, Nadler JT, Schwarz S, et al. 2018. A demonstration of the Collaborative Replication and Education Project: replication attempts of the red-romance effect. PsyArXiv, June 22. **https://doi.org/10.31234/osf.io/chax8**

Whitcomb D, Battaly H, Baehr J, Howard-Snyder D. 2017. Intellectual humility: owning our limitations. *Philos. Phenomenol. Res.* 94(3):509–39

Wicherts JM, Bakker M, Molenaar D. 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE* 6(11):e26828

Wiktop G. 2020. Systematizing Confidence in Open Research and Evidence (SCORE). *Defense Advanced Research Projects Agency.* **https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence**

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3(1):160018

Wilson BM, Harris CR, Wixted JT. 2020. Science is not a signal detection problem. *PNAS* 117(11):5559–67

Wilson BM, Wixted JT. 2018. The prior odds of testing a true effect in cognitive and social psychology. *Adv. Methods Pract. Psychol. Sci.* 1(2):186–97

Wintle B, Mody F, Smith E, Hanea A, Wilkinson DP, et al. 2021. Predicting and reasoning about replicability using structured groups. MetaArXiv, May 4. **https://doi.org/10.31222/osf.io/vtpmb**

Yang Y, Youyou W, Uzzi B. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *PNAS* 117(20):10762–68

Yarkoni T. 2019. The generalizability crisis. PsyArXiv, Nov. 22. **https://doi.org/10.31234/osf.io/jqw35**

Yong E. 2012. A failed replication draws a scathing personal attack from a psychology professor. *National Geographic*, March 10. **https://www.nationalgeographic.com/science/phenomena/2012/03/10/failed-replication-bargh-psychology-study-doyen/**

Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018. Improving social and behavioral science by making replication mainstream: a response to commentaries. *Behav. Brain Sci.* 41:e157