Replicability, Robustness, and Reproducibility in Psychological Science

Brian A. Nosek	Center for Open Science; University of Virginia
Tom E. Hardwicke	University of Amsterdam
Hannah Moshontz	University of Wisconsin-Madison
Aurélien Allard	University of California, Davis
Katherine S. Corker	Grand Valley State University
Anna Dreber	Stockholm School of Economics; University of Innsbruck
Fiona Fidler	University of Melbourne
Joe Hilgard	Illinois State University
Melissa Kline Struhl	Center for Open Science
Michèle Nuijten	Meta-Research Center; Tilburg University
Julia Rohrer	Leipzig University
Felipe Romero	University of Groningen
Anne Scheel	Eindhoven University of Technology
Laura Scherer	University of Colorado Denver - Anschutz Medical Campus
Felix Schönbrodt	Ludwig-Maximilians-Universität München, LMU Open Science Center
Simine Vazire	University of Melbourne

In press at the *Annual Review of Psychology*

Final version Completed: April 6, 2021

Authors' note: 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. This work was supported by grants to B.A.N. from Arnold Ventures, 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 programme under the Marie Skłodowska-Curie grant agreement No. 841188. We thank Adam Gill for assistance creating Figures 1, 2, and S2. Data, materials, and code are available at https://osf.io/7np92/. B.A.N. Drafted the outline and manuscript sections, collaborated with section leads on conceptualizing and drafting their components, coded replication study outcomes for Figure 1, coded journal policies for Figure 5. T.E.H. Drafted small sections of the manuscript (in 'What happens after replication?' and 'Evidence of change'). Collected and analyzed data for Figure 5 and Figure S1. Made suggestions and revisions to the manuscript prior to submission. H.M. Drafted small sections of the supplement ("Are behaviors changing?") and contributed content to sections of the manuscript ("Evidence of change" section). Compiled,

curated, and synthesized data for tables S4, S5, and S6. Contributed to curating data for Figure 1. A.A. Drafted small sections of the manuscript. Compiled, curated, and analyzed data for Figure 1. K.S.C. Drafted small sections of the manuscript. Made revisions to the manuscript. A.D. Drafted small sections of the manuscript. Compiled and analyzed data for Figure 2 and Figure S2. F.F. Drafted small sections of the manuscript. Made revisions to the manuscript. J.H. Drafted small sections of the manuscript (“Evidence of Change”) and supplement (“Retractions”). Analyzed data for these sections. Generated Figure 4. M.K.S. Drafted small sections of the manuscript. Made revisions to the manuscript. M.B.N. Drafted small sections of the manuscript. Made revisions to the manuscript. J.R. Drafted small sections of the manuscript. Made revisions to the manuscript. F.R. Drafted small sections of the manuscript and parts of supplement "What happens after replication?" Made revisions to the manuscript. A.S. Drafted small sections of the manuscript. Made revisions to the manuscript. L.S. Drafted small sections of the manuscript (“Evidence of change” section). Contributed to literature review appearing on pages 20-22, and coding appearing in Tables S4, S5, and Figure 4. Made suggestions and revisions to manuscript prior to submission. F.S. Drafted small sections of the manuscript. Collected and analyzed data for the section on “changes in job advertisements”. S.V. Drafted small sections of the manuscript. Made revisions to the manuscript.

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 understanding 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 understanding and observed surprising failures to replicate many published findings. Replication efforts highlighted sociocultural challenges, such as disincentives to conduct replications, framing of replication as personal attack rather than healthy scientific practice, and headwinds for replication contributing to self-correction. Nevertheless, innovation in doing and understanding replication, and its cousins, reproducibility and robustness, have positioned psychology to improve research practices and accelerate progress.

150 words

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

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 over-emphasis on statistical significance ($p < .05$), publication bias, inadequate statistical power, weak specification of theories and analysis plans, and lack of replication on the credibility of published findings (Cohen, 1973, 1994; Greenwald, 1975; Meehl, 1978; Rosenthal, 1979; Sterling, 1959). But those worries had little impact until conceptual and empirical contributions illustrated their potential ramifications for research credibility (Bakker et al., 2012b; Open Science Collaboration, 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 different analysis strategy. Reproducibility refers to testing the reliability of a prior finding using the same data and same analysis strategy (National Academies of Sciences, 2019). Each 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, then the same result should recur. Reproducibility tests can fail for two reasons. A *process reproducibility failure* occurs when the original analysis cannot be repeated because of unavailability of the data, code, information about the analysis to recreate the code, or necessary software or tools. An *outcome reproducibility failure* occurs when the reanalysis obtains a different result than reported originally. This can occur because of an error in either the original or 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 232 findings, and 18 of those only after deviating from the reported analysis in the paper (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 it 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 variation 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., 2019). 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 the 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 itself, 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, 2020; 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 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 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 would revise confidence in the original finding (Nosek & Errington, 2020). For example, if positive outcomes would increase confidence and expand generalizability, but negative outcomes would merely identify a potential boundary condition and 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 the position that distinctions such as direct versus conceptual replication are counterproductive (Machery, 2020; Nosek & Errington, 2020). This position guides this review but is not uncontested. For alternative perspectives about the value of replication and terminological distinctions of types of replications see Crandall & Sherman, 2016; LeBel et al., 2018; Schwarz & Strack, 2014; Simons, 2014; Stroebe & Strack, 2014; Wilson et al., 2020; 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 $p < .05$, despite persistent exhortations by methodologists. The desire for dichotomous simplicity occurs with replications too: “Did it 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 of conditions or effect size. 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 dichotomous assessment of replication outcomes, each of which provide some perspective and none of which are definitive. These include assessing whether the replication rejects the null hypothesis ($p < .05$) in the same direction as the original study (Camerer et al., 2018; Open Science Collaboration, 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 Science Collaboration, 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 subjective assessment of whether the findings are similar (Open Science Collaboration, 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 of whether a replication failed or succeeded.

As psychological theory and evidence matures, 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 size and cumulative evidence via meta-analysis or related approaches (Mathur & VanderWeele, 2020). Meta-analysis in replication research examines the average effect size, degree of uncertainty, and evidence for heterogeneity across samples, settings,

treatments, and outcomes (Hedges & Schauer, 2019; Landy et al., 2020). When heterogeneity is high, it can indicate that there are moderating influences to identify, test, and then improve theoretical understanding. Meta-analyses, however, are often undermined by publication bias favoring significant results and other threats to the quality of individual studies (Carter et al., 2019; Rothstein et al., 2005; Vosgerau et al., 2019). Averaging 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. New evidence is added reinforcing or reorganizing present understanding. The cycle repeats.

A Note on Validity

A finding can be reproducible, robust, replicable, and invalid at the same time. Credibility via reproducibility, robustness, and replicability does not guarantee that treatments worked as intended, measures assessed the outcomes of interest, or that interpretations correspond with the evidence produced. But conducting replications can help identify sources of invalidity if those sources of invalidity vary in the operationalizations of replications that are believed to be testing the same phenomenon. Deliberate efforts to root out invalidity via replications can also be productive. For example, an original finding might be 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 but without influencing other variables. Stated this way, the ordinariness and centrality of replication for advancing knowledge becomes clear. Many replications, in practice, are efforts to root out invalidity, either because of questionable research practices that undermine the statistical evidence or because of 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 might be lower than expected or desired have been available for decades. Cohen and others (Button et al., 2013; Cohen, 1973, 1992a; 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 ($p < .05$). But other evidence suggests that 90% or more of primary outcomes are statistically significant (Fanelli, 2010, 2012; Sterling et al., 2012; Sterling, 1959). 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 influence of publication bias in which negative results are

ignored (Greenwald, 1975; Rosenthal, 1979) or questionable research practices that are inflating reported effect sizes (John et al., 2012; Simmons et al., 2011). Despite broad recognition of the disjoint, there were no systematic efforts to test the credibility of the literature until this past decade.

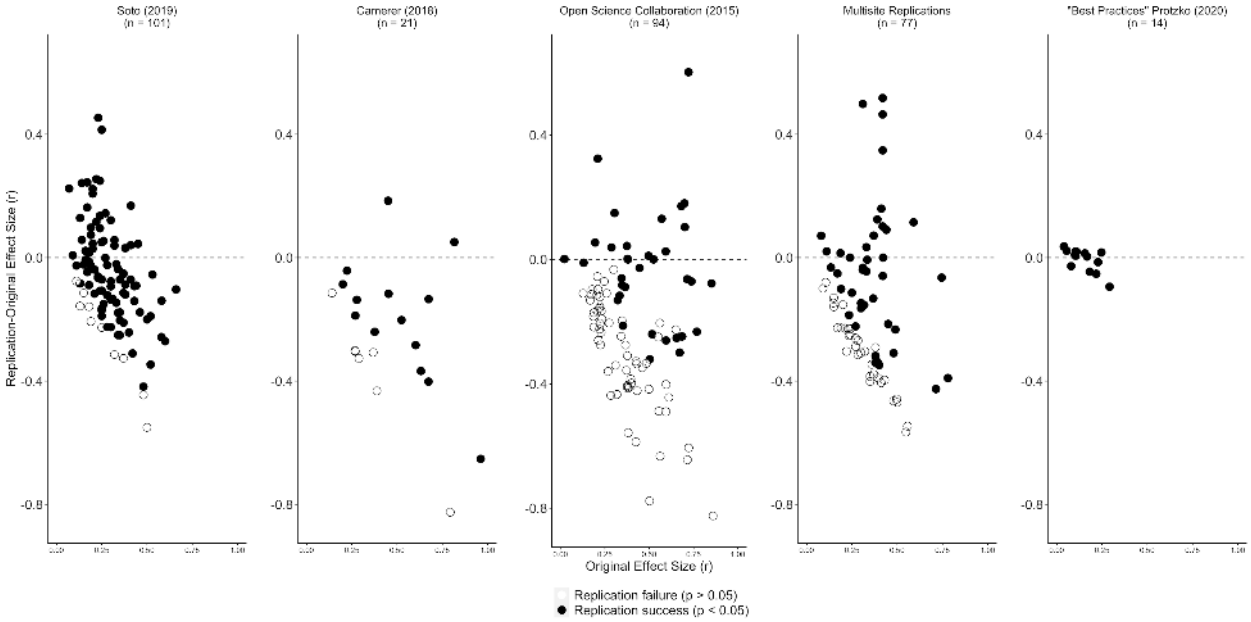
"How replicable is psychological research?" is unlikely to ever be answered with confidence, as "psychological research" is large, changing constantly, and has ill-defined boundaries. A large enough random selection of studies to make a precise and meaningful estimate exceeds feasibility. But progress can be made by benchmarking replicability of samples of findings against expectations of their credibility. For example, if an individual finding is regularly cited and used as the basis for supporting theory, then there exists an implicit or explicit presumption that it 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 increases as 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 empirical input for these discussions, we summarize recent evidence concerning replicability in psychology. We gathered two types of prominent replication studies conducted during the last decade: [1] "Systematic replications" were replication efforts that defined a sampling frame and conducted replications of as many studies in that sampling frame as possible to minimize selection biases, and [2] "Multi-site replications" conducted the same replication protocol of prominent findings in a variety of samples and settings to obtain highly precise effect size estimates and estimate 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.

For "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 the original studies. Camerer and colleagues (2018) replicated 21 social science experiments systematically selected from *Nature* and *Science* papers published between 2010 and 2015; 62% achieved significance in the same direction with effect sizes 50% as large as the original studies. Open Science Collaboration (2015) replicated 100 findings from 2008 issues of three psychology journals; 36% achieved significance in the same direction with effect sizes 49% as large as the original studies. "Multi-site replications" include the series titled "Many Labs" (Ebersole, Atherton, et al., 2016; Ebersole et al., 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; E.-J. 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 Consortium, 2020; Mccarthy et al., 2020; McCarthy et al., 2018; Moran et al., 2020; Schweinsberg et al., 2016).

Collectively (*n* = 77), 56% of multisite replications reported statistically significant evidence in the same direction with effect sizes 53% as large as the original studies (Figure 1).

Figure 1. Replication outcomes for three systematic replication studies (Soto [2019], Camerer [2018], Open Science Collaboration [2015]), multisite replication studies, and a prospective "best practice" replication study (Protzko, 2020). Values above zero indicate 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.



Combining across all replications (*n* = 307), 64% reported statistically significant evidence in the same direction with effect sizes 68% as large as the original studies. Moreover, the sample size for replication studies was on average 15.1 times the size of the original studies (Median = 2.8; STD = 32.8) eliciting more precise estimates of effect size and heterogeneity, and leading to a relatively generous definition of "success" with high power to detect a significant effect in the same direction as the original in the binary categorization of the replication outcome in these figures.

We cannot estimate the replicability rate or effect size overestimation of 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 evidence for how these systematic and multisite replications underestimated replicability.

What replicates and what doesn't?

Some replications produce evidence consistent with the original studies—others do not. Why is that? Knowledge about the correlates and potential causes could help advance

interventions to increase replicability. We discuss three overlapping classes of correlates; theoretical maturity, features of the original studies, and features of the replication studies.

Theory

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, but 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 aspect of theoretical maturity is a good understanding 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, 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. Insufficient detail makes it more 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 despite both being well-conducted when our understanding is not yet mature enough to anticipate the consequences of seemingly irrelevant factors in the sample, setting, interventions, and outcome measures (Nosek & Errington, 2020; 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., 2018a). Van Bavel and colleagues (2016) observed that judges' ratings of the context sensitivity of phenomena included in Open Science Collaboration (2015) were negatively associated with replication success (r = -0.23). However, Inbar (2016) observed that the correlation did not hold within social (r = -.08) and cognitive (r = -.04) subdisciplines and thus 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 of presumed context sensitivity failing to account for replication failures or weaker effect sizes than original studies when examined directly (Ebersole, Atherton, et al., 2016; Ebersole et al., 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 circumstances in which the replicability of a finding is

demonstrated to be contingent on unconsidered factors in the operationalization will be productive for advancing investigations of 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 for 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 Science Collaboration (2015) to be just 8.8% (range 0.7% to 66%). Relatedly, Open Science Collaboration (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 Science Collaboration (2015) reported exploratory analyses that lower p-values in original studies were associated with higher likelihood of replication success (r = -0.33). In a literature with relatively small sample sizes and publication bias favoring positive results, large observed effects can be a sign of overestimated or false positive effects rather than 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 more 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 with just the information provided in original papers and supplements. In no case were they able to create a comprehensive protocol without asking clarifying questions of 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 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; Wilkinson et al., 2016), following reporting standards such as JARS (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 the process and context of obtaining original findings transparently may reduce the replicability of reported findings. For example, researchers are more likely to publish positive findings than negative findings (Franco et al., 2014; Franco et al., 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,

2008; Ioannidis, 2005). Conducting multiple studies and only reporting a subset that achieve statistical significance will inevitably reduce replicability if research includes any false hypotheses (Nuijten, van Assen, et al., 2015; Schimmack, 2012). And, 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., 2012b; Götz et al., 2020; Nelson et al., 2018; Simmons et al., 2011).

Preregistration of studies in a registry ensures that studies are, in principle, discoverable even if they are not reported in published articles. Increasing transparency and discoverability of all studies will improve accuracy of meta-analyses and 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 same 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 with 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 in which 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 are 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.

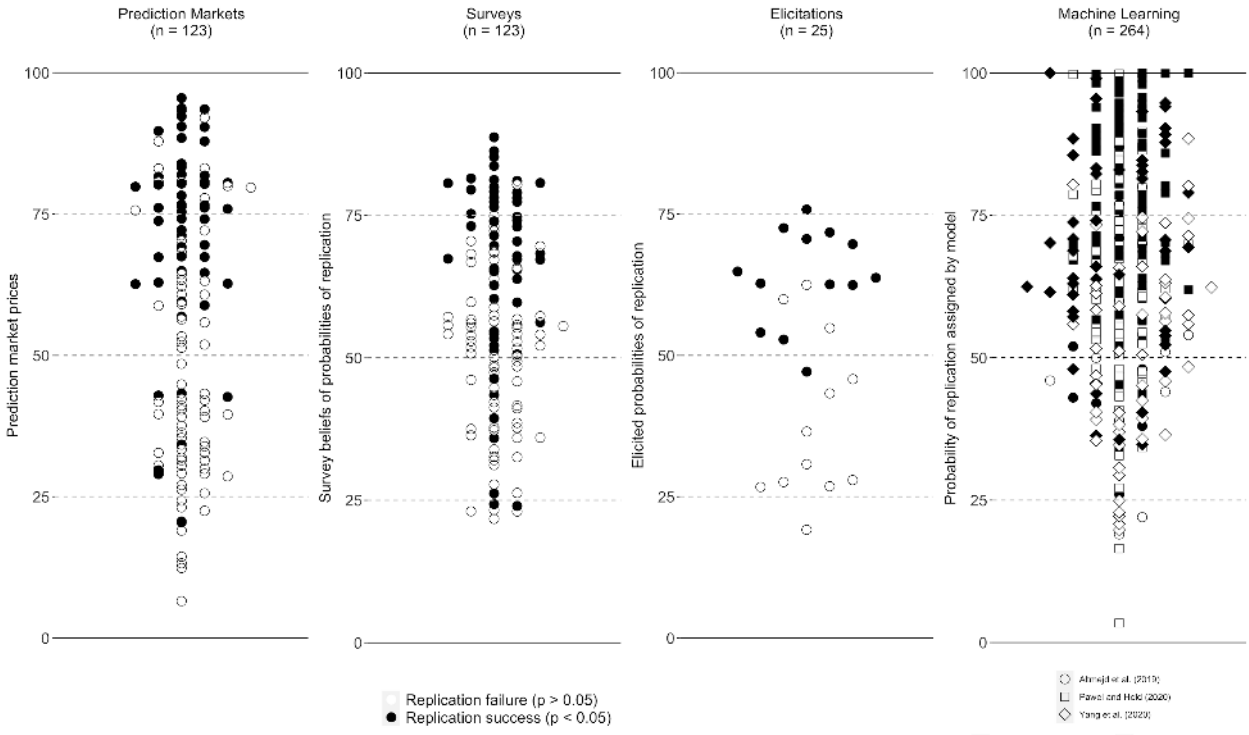
Further, specific cases that have been empirically examined do not support the conclusion that the replication failures occurred because of incompetence or of failing to implement and conduct the study appropriately. Gilbert and colleagues (2016) speculated that Open Science Collaboration (2015) failed to replicate many studies because the replication teams did not conduct the experiments with sufficient fidelity to the originals and highlighted original authors' lack of endorsement of some protocols as supporting evidence (but see Anderson et al., 2016; Nosek & Gilbert, 2017). And, they 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 $N = 1279$) using the same replication protocols in one condition and protocols revised following formal peer review by experts in another condition. Formal expert review produced little to no increase in effect sizes compared with 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 features of original studies and findings. It might also foster development of faster and cheaper indicators of replicability to guide attention and resource allocation towards 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--suggest that replication outcomes are predictable. Surveys present brief descriptions of original studies and findings and then average individual estimates about likelihood of successful replication. Structured elicitations engage small groups of individuals to make private initial judgments, and then compare and discuss estimates and share information with group members before making a second final private judgment (Hanea et al., 2017). Structured protocols such as IDEA (Investigate Discuss Estimate Aggregate) rely on mathematical aggregation of group members' final judgments, rather than forcing behavioural consensus in the way traditional delphi groups do. Prediction markets have participants bet on whether the studies 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 assigns the replication outcome to be successful (Dreber et al., 2015).

Figure 2. Predictions of replication outcomes across four methods: Surveys, Elicitations, Prediction Markets, 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.



In all approaches, aggregating across projects, predictions were positively correlated with observed replication success (*r*'s = 0.52 [prediction markets], 0.48 [surveys], 0.75 [elicitations]; Figure 2). Using a dichotomous criterion of prices above 50 anticipating replication success and below 50 anticipating replication failure in the prediction markets, 88 of 123 (72%) and 79 of 123 (64%) in similar survey formats were predicted accurately. Prediction markets on replication outcomes based on effect size have so far 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 the 123 prediction-replication pairs, lay people predicted replication success with 59% accuracy, which increased to 67% when also receiving 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 after training predictive models and then doing out-of-sample tests. The results reinforce the conclusion that statistical properties like sample sizes, *p*-values and effect sizes of the original studies, and 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 perform 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 to inform evaluation, resource allocation, and identification of gaps and strengths in the empirical evidence for theoretical models and findings. A third study used a different type of forecasting approach using the original studies' information and the replication studies' sample size only (Pawel &

Held, 2020). For the comparable samples, the forecasts from the tested statistical methods performed as well as or worse than the prediction markets.

What degree of replicability should be expected?

Non-replicable findings are a risk factor for research progress, 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 non-replicability should decline with maturation of a research topic, but a healthy, theoretically-generative research enterprise will include non-replicable findings. Simultaneously, a healthy, theoretically-generative research enterprise will constantly be 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 will 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 includes increasing numbers 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 includes setting stricter inference criteria (Benjamin et al., 2018; Lakens et al., 2018); taking precautions against *p*-hacking, HARKing, selective reporting by employing preregistration and transparency; and taking alternative explanations seriously by conducting 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 generality of the findings (Simons et al., 2017), and calibrating certainty based on the extent to which the statistical inferences could have been 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 conduct follow-up research. Evaluation is facilitated by maximizing transparency of the research process, including sharing methods, materials, procedures, and data, reporting the timing of decisions and any data-dependency in analyses (Lakens, 2019; Nosek et al., 2019), and 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 planning new research should include the expectation of smaller effect sizes and planning for larger samples to detect them (Funder & Ozer, 2019; Perugini et al., 2014).

Initial evidence suggests that behavioral changes of preregistration, large samples, and sharing of research materials in original studies are associated with high replicability. Protzko and colleagues (2020) implemented these behaviors in a “best practices” 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 with original findings from other replication projects. But, those original effect sizes appear to be credible: Replication effect sizes were 97% as large as the original studies suggesting that high replicability is achievable. This study does not, however, provide causal evidence of specific practices increasing 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 for proposed studies based on the importance of the question and the quality of the methodology before knowing the outcomes (Chambers, 2019). Registered Reports provides a structural solution to support selecting good research questions, using appropriately severe methods and procedures, preregistered planned analyses, and presenting work transparently for others to provide critique. Using Registered Reports is also associated with high rates of sharing data and materials and higher ratings of quality and rigor than comparison papers (Soderberg et al., 2020), and much higher proportion of reporting null results (Scheel et al., 2020). As 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 clear 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

Knowledge of the underlying causes of non-replicability and solutions for improvement are not sufficient on their own to improve replicability. The problems and solutions for low powered research and misuse of null hypothesis significance testing have been understood since before Jacob Cohen started writing about them in the 1960s (Cohen, 1962). Indeed, reflecting on the lack of increase of sample size and power 30 years later he wrote “I have learned, but not easily, that things take time” (Cohen, 1992b, p. 1311), and “we must finally rely, as have the older sciences, on replication” (Cohen, 1994, p. 1002). Psychological science is a complex system with structural constraints, social context, 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 impede the ability and opportunity to do it (Hardwicke et al., 2020).

Social and Structural Context

Academic science occurs in a complex system of policies, norms, and reward systems that are shaped by decisions about which research is funded, which research gets published, and which researchers get jobs and promotions. The system of policies, incentives, and norms is strong enough that researchers may value behaviors that improve replicability and know how to do those behaviors, but still not do them 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., 2012a), 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). With substantial discretion on which studies and analyses are reported, researchers have both motivation and opportunity to engage intentionally or unintentionally in behaviors that improve publishability at the cost of credibility.

These trends affect all areas of science but 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 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 another's theory for novel research is an impediment to 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 potential with it. 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 researchers' self-interest to run many, underpowered studies than fewer, well-powered ones (see also Gervais et al., 2015; Tiokhin & Derex, 2019). The combination of these is combustible--a literature filled with novel contributions selectively reported from many underpowered studies without replications to assess credibility or improve precision. The lack of replication also means that there are no costs for publishing false positives--reputational or otherwise--thus reinforcing the singular emphasis on novelty.

Smaldino and McElreath (2016) incorporated some of these structural incentives into a dynamic model of the research community conducting research and publishing results and 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 is insufficient to alter the dysfunctional process, emphasizing that structural change is also necessary for effective reform.

Imbuing scholarly debates about evidence with negative attributions about character and intentions interferes with treating replication as ordinary, good practice (Meyer & Chabris, 2014; Yong, 2012). This creates a fraught social environment that discourages researchers from skeptical and critical inquiry about existing findings. This dysfunctional social culture might be sustained by defining researcher identities 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 trade-off between a researcher who conducts boring but reproducible research versus one who conducts exciting but not reproducible research, raters consistently and strongly favor the former (Ebersole, Axt, et al., 2016). Moreover, researchers who take seriously an independent failure to replicate by acknowledging it or conducting follow-up research receive reputational enhancement, even more so than researchers whose findings replicate successfully (Ebersole, Axt, et al., 2016; Fetterman & Sassenberg, 2015). This suggests that the challenges are rooted in the social and structural features of science, not in the minds of practicing scientists. Highlighting the consequences of some of the 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 researcher reputations to their findings, people like to be right (Kunda, 1990). But, scientific progress is made by identifying where current understanding is wrong and generating new ideas that might make it less wrong. So, part of 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 intellectual humility more generally (Ebersole, Axt, et al., 2016; Leary et al., 2017; Whitcomb et al., 2017). Intellectual humility may be relevant to a variety of reasoning biases that can interfere with pursuit of truth including confirmation bias in which researchers might selectively attend to or create conditions that are consistent with their existing positions (Nickerson, 1998), hindsight bias in which researchers might revise their theoretical “predictions" about what they would have expected of a replication design after observing the outcomes (Christensen-Szalanski & Willham, 1991; Kerr, 1998), and outcome bias in which researchers evaluate the quality of replication designs 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 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 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 comprise a decentralized behavior change strategy that is transforming how research is done, reported, and evaluated.

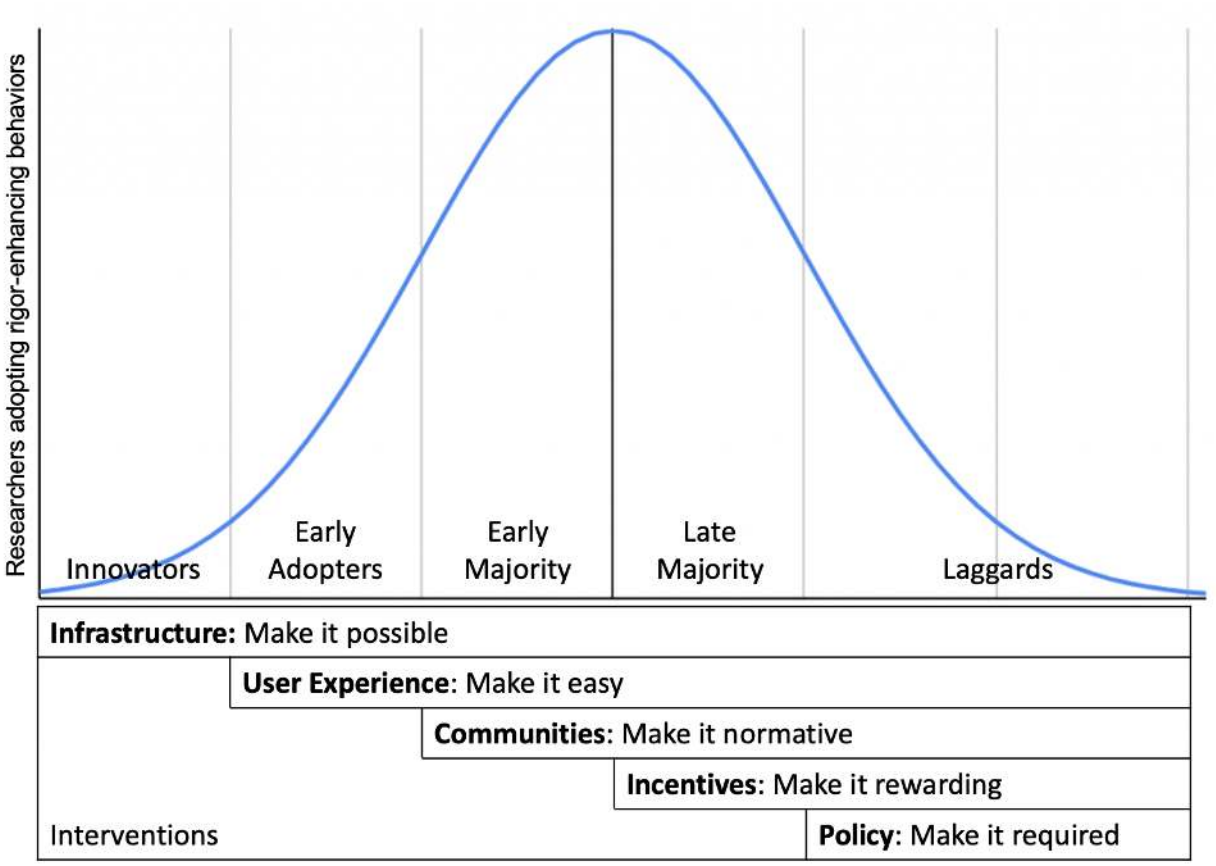
Strategy

The culture change movement has many stakeholders making independent decisions about whether and how to change their policies and practices to improve replicability. There also has 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 missions to promote culture change toward rigor, transparency, and replicability. COS's culture change strategy, based on Rogers' diffusion model (Figure 3), is a reasonable account of the broader, decentralized actions by many groups that are fostering culture change (Nosek, 2019).

The model extends Rogers' (2003) theoretical work on diffusion of innovations for how new technologies are first used by innovators and early adopters and then gain mainstream acceptance. It rests on a few key principles: culture and behavior change unfold over time, motivations differ across people and circumstances between the introduction and 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.

According to the extension of the diffusion model, for innovators who are motivated by trying and testing new behaviors, providing infrastructure and tools that *make it possible* to do the behavior can be sufficient for 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 do the behaviors. Those early adopters are critical for achieving mainstream adoption based on their direct, grassroots social influence of peers and the indirect visibility of their behaviors more generally to *make it normative* to do the behaviors. Bottom-up behavior change will eventually stall if there is not stakeholder support to shift incentives to *make it desirable* to do the behaviors. And, even incentives may not be sufficient to unseat behaviors that are sustained by structural factors. Policy changes adapt the structure and *make it required* or part of the system to do the behaviors.

Figure 3. Interdependent interventions for effective culture change extending Rogers' (2003) diffusion model.



The model's five levels of intervention are highly interdependent, each necessary, and none sufficient alone for effective culture and behavior change. For example, a policy intervention that does not have quality infrastructure or normative support is likely to fail to meet its intentions because meeting the policy is difficult and not valued, turning the policy from promoting good practice into imposing unwelcome bureaucratic burden.

Evidence of change

Behaviors that may directly or indirectly improve replicability, or the ability to assess replicability, include increasing sample size, preregistration, improving rigor and transparency, sharing materials and primary data, conducting replications, and enhancement of 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; Soderberg, 2018) and AsPredicted; error detection and correction such as statcheck (Epskamp & Nuijten, 2018) and GRIM (Granularity Related Inconsistent Means; Brown & Heathers, 2017); grassroots communities promoting new norms such as the *Society for Improving Psychological Science, Open Science Communities* (Armeni et al., 2020), national reproducibility networks (Munafò et al., 2020); large-scale collaboration to increase sample size and replication efforts such as Psychological Science Accelerator (Moshontz et al., 2018) and ManyBabies (Byers-Heinlein et al., 2020); increasing visibility of behaviors to shift norms such as badges for open practices (Kidwell et al., 2016); altering incentives for publishing away from

positive, novel, tidy results 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 such as TOP Guidelines (Nosek et al., 2015).

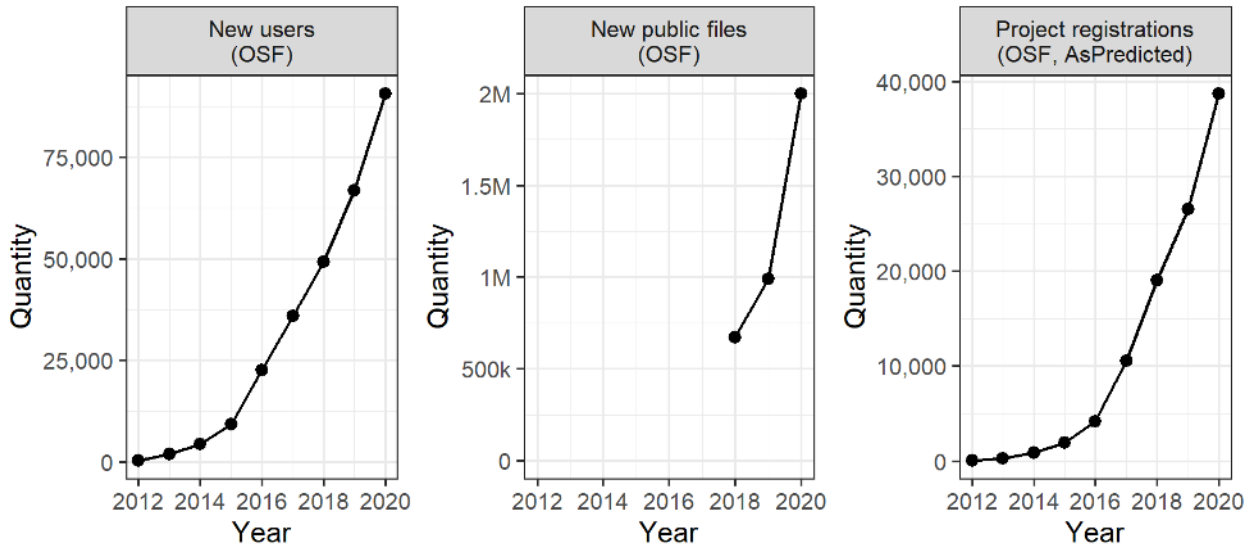
Most available survey data suggests that psychologists and other social-behavioral researchers acknowledge engaging in questionable research practices that could interfere with replicability (John et al., 2012). Table S5 summarizes 14 surveys of QRPs (Total *N* = 7,887) with, for example, 9% to 43% of researchers acknowledging failing to report all study outcomes and 25% to 62% acknowledging 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.

Seven surveys of researchers (Total *N* = 4,737; Table S3) and four audit studies (Total *N* articles = 1,100; Table S4) assessed behaviors like sharing data and materials or preregistration among psychological scientists. Surveys observed between 27% and 60% reporting having shared data and between 27% and 57% reporting having preregistered a study. Audit studies observed high variation in data sharing of 0% to 65%, likely based on their sampling strategy, and only one study assessed preregistration and observed a rate of 3%.

The audit studies suggest that self-reported behaviors have not yet translated in high numbers into the published literature itself with, for example, 2% of a random sample of psychology studies published between 2014 and 2017 having shared data and 3% having a preregistered study (Hardwicke et al., 2020). The discrepancy between audits and self-report is likely a function of multiple factors including when the surveys and audits were conducted, possible overreporting of the behavior in surveys, the time lag between doing the behavior and it appearing in a published article, and that surveys tend to ask about conducting the behavior once whereas individual researchers conduct many 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 pre-registered a study or posted data or code online, and observed about 20% having shared data or code and about 8% having preregistered in 2011 with those numbers rising to 51% and 44% by 2017. Supporting that self-reported evidence, usage of services like OSF and AsPredicted for preregistration and sharing data and materials have grown exponentially (Figure 4). Both services are available to anyone, but a substantial portion of their user bases are from psychology and allied fields. A 2019 analysis of all faculty from 69 psychology departments (*N* = 1,987) indicated that 35% had OSF accounts with heaviest representation in social (57%), quantitative (48%), and cognitive (42%), and lightest representation in clinical (19%) and education and health (17%; Nosek, 2019).

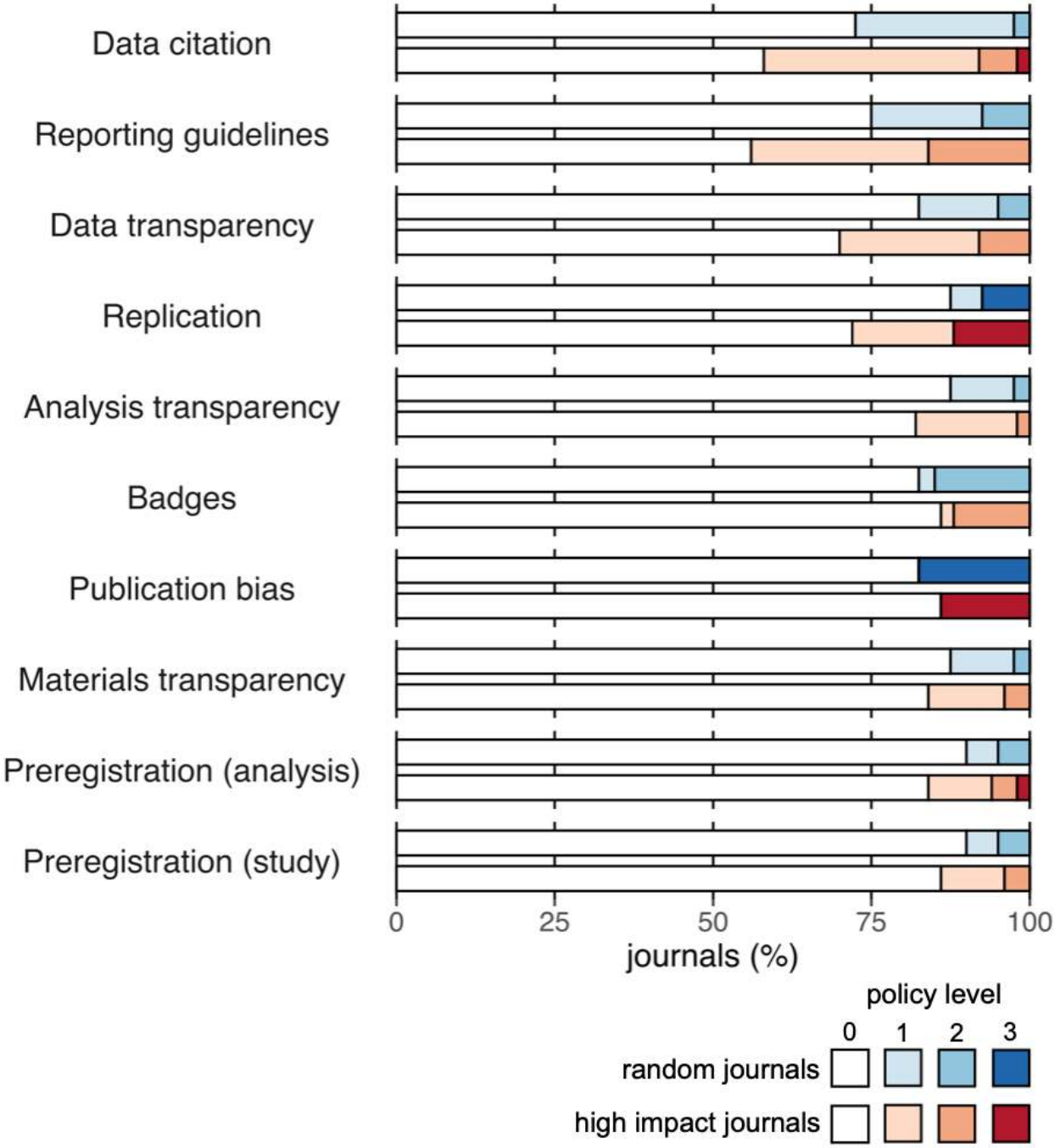
Figure 4. Yearly counts of users, sharing of files (research data, materials, code), and registration of studies on OSF and AsPredicted, two popular services for psychologists and allied disciplines. Data for new public files (sharing) prior to 2018 not available.



Contrasting the evidence of behavior change, in supplementary files we summarize five investigations of sample size over time and do not observe compelling evidence of change from 1977 to 2017, and studies of statistical reporting errors over time suggest relative stability from 1985 to 2013 (Bakker & Wicherts, 2011; Nuijten, Hartgerink, et al., 2015). Both require updated investigations for the latter half of the decade. Also, in the supplement 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 towards detection and correction of misconduct, faulty research, and honest errors (Marcus & Oransky, 2018).

Evidence of individual behavior change is complemented by adoption of normative, incentive, and policy interventions by journals and other stakeholder groups. Idiosyncratic actions by stakeholders have given prominence to replicability such as the Netherlands Organization for Scientific Research, National Science Foundation, and DARPA issuing calls for replication research proposals (Baker, 2016; Cook, 2016; Root, 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 interest in incorporating such factors into hiring and promotion (*TOP Resources - Evidence and Practices*, 2016). We conducted two systematic inquiries to assess the current status of journal policies and psychology department hiring practices.

Figure 5. Adoption of 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 (1, 2, or 3) from the most at the top to the least at the bottom.



The Transparency and Openness Promotion (TOP) guidelines are a set of 10 policy standards related to transparency and reproducibility, each with 3 levels of increasing stringency (Nosek et al., 2015; Table S7). We assessed adoption of TOP-compliant policies in a random sample of psychology journals (n = 40; *random journals*) and the 5 journals with the highest impact factor from each of 10 psychology subfields (n = 50; *high impact journals*). Methodological details are available in the supplement. As illustrated in Figure 5, for each of the ten standards, the substantial majority of journals had not adopted TOP-compliant policies (i.e., level 0; range 56-90%, median = 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 (17% and 15% respectively). Combining samples, TOP-compliant policies were most common for citing data sources (36%) and using reporting guidelines (36%), and were least common for preregistration of studies (12%) and analysis plans (13%). These findings suggest modest adoption of replicability-related policies among psychology journals. Notably, psychology's largest publisher, APA journals, has indicated an intention to move all of its core journals to at least TOP level 1 across eight standards by the end of 2021 (Center for Open Science, 2020).

We also examined whether research institutions are explicitly communicating expectations for replicability and transparency in their job advertisements. We analyzed all academic job offers in psychology from the German platform academics.de from February 2017 to December 2020 (*n*=1626). 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, the remainder (*n*=12) 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 (2017 and 2018: 1.0%; 2019: 2.0%; 2020: 3.8%).

There is both substantial evidence of new behaviors that may increase 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 leading change among stakeholder groups with department and institutional practices for hiring and promotion showing less 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., 2020; Zwaan et al., 2018a). 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 promoting 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 are necessary to improve 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? An earlier section 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 reform 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 having unintended negative consequences that outweigh the intended benefits? A healthy metascience and culture change movement will be constantly evaluating its 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 experiences 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.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., … Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLOS ONE*, *14*(12), e0225826. https://doi.org/10.1371/journal.pone.0225826

Anderson, C. J., Bahnik, t pan, Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., … Zuni, K. (2016). Response to Comment on “Estimating the reproducibility of psychological science.” *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad9163

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative Dissonance in Science: Results from a National Survey of U.S. Scientists. *Journal of Empirical Research on Human Research Ethics*, *2*(4), 3–14. https://doi.org/10.1525/jer.2007.2.4.3

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3. https://doi.org/10/gctpzj

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V. E., Heunis, S., Koh, W. Q., Masselink, M., Moran, N., Baoill, A. Ó., Sarafoglou, A.,

Schettino, A., Schwamm, H., Sjoerds, Z., Teperek, M., Akker, O. van den, Veer, A. van ’t, & Zurita-Milla, R. (2020). *Towards wide-scale adoption of open science practices: The role of open science communities*. MetaArXiv. https://doi.org/10.31222/osf.io/7gct9

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*. https://doi.org/10.1037/met0000365

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

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012a). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012b). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569. https://doi.org/10/fw72jd

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

Baumeister, R. F., & Vohs, K. D. (2016). Misguided Effort With Elusive Implications.

Perspectives on Psychological Science, *11*(4), 574–575. https://doi.org/10/gf5srq

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., … Wollbrant, C. E. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson,

S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144.

Center for Open Science. (2020, November 10). *APA Joins as New Signatory to TOP Guidelines*. https://www.cos.io/about/news/apa-joins-as-new-signatory-to-top-guidelines

Cesario, J. (2014). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Chambers, C. (2019). What's next for Registered Reports? *Nature*, *573*(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., … Yong, J. C. (2016). Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. https://doi.org/10.1177/1745691616664694

Christensen, G., Wang, Z., Paluck, E. L., Swanson, N., Birke, D. J., Miguel, E., & Littman, R. (2019). *Open Science Practices are on the Rise: The State of Social Science (3S)*

Survey. https://doi.org/10.31222/osf.io/5rksu

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*(1), 147–168.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1973). Brief Notes: Statistical power analysis and research results. *American Educational Research Journal*, *10*(3), 225–229.

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*(1), 155.

Cohen, J. (1992b). Things I have learned (so far). *Annual Convention of the American Psychological Association, 98th, Aug, 1990, Boston, MA, US; Presented at the Aforementioned Conference.*

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Colling, L. J., Sz\Hucs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., & Schroeder, P. A. (2020). Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003). *Advances in Methods and Practices in Psychological Science*, 2515245920903079.

Cook, F. L. (2016, September 20). *Dear Colleague Letter: Robust and Reliable Research in the Social, Behavioral, and Economic Sciences (nsf16137) | NSF - National Science Foundation*. https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Crisp, R. J., Miles, E., & Husnu, S. (2014). *Support for the replicability of imagined contact effects.*

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., De Cristofaro, V., de Vries, L., Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., … Zinkernagel, A. (2021). A Multilab Replication of the Ego Depletion Effect. *Social Psychological and Personality Science*, *12*(1), 14–24. https://doi.org/10/ggtpft

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

Dijksterhuis, A. (2018). Reflection on the professor-priming replication report. *Perspectives on Psychological Science*, *13*(2), 295–296. https://doi.org/10.1177/1745691618755705

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton University Press.

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., Chung, L. Y., Hermann, A. D., Joy-Gaba, J. A., Line, M. J., Rule, N. O., Sacco, D. F., Vaughn, L. A., & Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology*, 3.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' Reputations Are Based on Getting

It Right, Not Being Right. *PLOS Biology*, *14*(5), e1002460. https://doi.org/10.1371/journal.pbio.1002460

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Ellemers, N., Fiske, S. T., Abele, A. E., Koch, A., & Yzerbyt, V. (2020). Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *Proceedings of the National Academy of Sciences*, *117*(14), 7561–7567. https://doi.org/10.1073/pnas.1906720117

Epskamp, S., & Nuijten, M. B. (2018). *Statcheck: Extract statistics from articles and recompute p values. R package version 1.3.1.* https://CRAN.R-project.org/package=statcheck

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). *Challenges for Assessing Reproducibility and Replicability Across the Research Lifecycle in Preclinical Cancer Biology*.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, *11*(2), e0149794. https://doi.org/10.1371/journal.pone.0149794

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Feest, U. (2019). Why Replication Is Overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10/gg4357

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). *Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes.*

Fetterman, A. K., & Sassenberg, K. (2015). The Reputational Consequences of Failed Replications and Wrongness Admission among Scientists. *PLOS ONE*, *10*(12), e0143723. https://doi.org/10.1371/journal.pone.0143723

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, *75*, 102117. https://doi.org/10.1016/j.joep.2018.10.009

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Franco, Annie, Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments: Evidence From a Study Registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

https://doi.org/10.1177/2515245919847202

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

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. *Department of Statistics, Columbia University*, *348*.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309. https://doi.org/10/fgjg7z

Gervais, W. M., Jewell, J. A., Najle, M. B., & Ng, B. K. L. (2015). A Powerful Nudge? Presenting Calculable Consequences of Underpowered Research Shifts Incentives Toward Adequately Powered Designs. *Social Psychological and Personality Science*, *6*(7), 847–854. https://doi.org/10.1177/1948550615584199

Ghelfi, E., Christopherson, C. D., Urry, H. L., Lenne, R. L., Legate, N., Ann Fischer, M., Wagemans, F. M. A., Wiggins, B., Barrett, T., Bornstein, M., de Haan, B., Guberman, J., Issa, N., Kim, J., Na, E., O’Brien, J., Paulk, A., Peck, T., Sashihara, M., … Sullivan, D. (2020). Reexamining the Effect of Gustatory Disgust on Moral Judgment: A Multilab Direct Replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science*, *3*(1), 3–23. https://doi.org/10.1177/2515245919881152

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science.” *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological*

Science, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Giner-Sorolla, R. (2019). From crisis of evidence to a “crisis” of relevance? Incentive-based answers for social psychology’s perennial relevance worries. *European Review of Social Psychology*, *30*(1), 1–38. https://doi.org/10.1080/10463283.2018.1542902

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)*. https://doi.org/10.23668/psycharchives.3010

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., & Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, *7*(7), 200566. https://doi.org/10.1098/rsos.200566

Götz, M., O’Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2020). The “Goldilocks Zone”:(Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10/bfwsfj

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hanea, A. M., McBride, M. F., Burgman, M. A., Wintle, B. C., Fidler, F., Flander, L., Twardy, C. R., Manning, B., & Mascaro, S. (2017). I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *International Journal of Forecasting*, *33*(1), 267–279.

https://doi.org/10.1016/j.ijforecast.2016.02.008

Hardwicke, Tom E., Bohn, M., MacDonald, K. E., Hembacher, E., Nuijten, M. B., Peloquin, B., deMayo, B., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society Open Science*, *8*(1). https://doi.org/10.1098/rsos.201494

Hardwicke, Tom E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hardwicke, Tom E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104

Hardwicke, Tom Elis, Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M., & Ioannidis, john. (2020). *Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017)* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/9sz2y

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557. https://doi.org/10/ggbnzr

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285.

Hughes, B. M. (2018). *Psychology in crisis*. Palgrave/Macmillan Education.

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology

may be spurious. *Proceedings of the National Academy of Sciences*, *113*(34), E4933–E4934. https://doi.org/10.1073/pnas.1608676113

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640–648. https://doi.org/10/cst2h8

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549.

Isager, P. M., Aert, R. C. M. van, Bahník, Š., Brandt, M., DeSoto, K. A., Giner-Sorolla, R., Krueger, J., Perugini, M., Ropovik, I., Veer, A. van 't, Vranka, M. A., & Lakens, D. (2020). *Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection.* MetaArXiv. https://doi.org/10.31222/osf.io/2gurz

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10/f33h6z

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*(9), 723.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456.

https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., & Crawford, J. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. https://doi.org/10/ghwq2w

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis*. https://doi.org/10.31234/osf.io/jbh4w

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., & Bradford, D. E. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10/gcz8f3

Landy, J. F., Jia, M. (Liam), Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how

design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10/f96wsf

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402.

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. (2018). Self-Esteem, Self-Disclosure, Self-Expression, and Connection on Facebook: A Collaborative Replication Meta-Analysis. *Psi Chi Journal of Psychological Research*, *23*(2), 98–109. https://doi.org/10.24839/2325-7342.JN23.2.98

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. https://doi.org/10.31234/osf.io/6btc3

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the History of Economic Thought and Methodology* (Vol. 36, pp. 129–146). Emerald Publishing Limited. https://doi.org/10.1108/S0743-41542018000036B009

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*(1), 358. https://doi.org/10.1038/s41467-019-14203-0

Maassen, E., Assen, M. A. L. M. van, Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10/gg2cz8

Machery, E. (2020). What is a Replication? *Philosophy of Science*. https://doi.org/10.1086/709701

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10/ghwq2t

Marcus, A., & Oransky, I. (2018, February 14). *Meet the 'data thugs' out to expose shoddy and questionable research*. Science | AAAS. 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*. https://www.wired.com/story/tech-firms-hire-red-teams-scientists-should-too/

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 1145–1166. https://doi.org/10/ghwq2s

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147. https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487. https://doi.org/10/f7qwxd

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Mccarthy, R., Gervais, W., Aczel, B., Al-Kire, R., Baraldo, S., Baruh, L., Basch, C., Baumert, A., Behler, A., Bettencourt, A., Bitar, A., Bouxom, H., Buck, A., Cemalcilar, Z., Chekroun, P., Chen, J., Díaz, Á., Ducham, A., Edlund, J., & Zogmaister, C. (2020). A Multi-Site Collaborative Study of the Hostile Priming Effect. *Collabra Psychology*.

McCarthy, R. J., Hartnett, J. L., Heider, J. D., Scherer, C. R., Wood, S. E., Nichols, A. L., Edlund, J. E., & Walker, W. R. (2018). An Investigation of Abstract Construal on Impression Formation: A Multi-Lab Replication of McCarthy and Skowronski (2011). *International Review of Social Psychology*, *31*(1), 15. https://doi.org/10.5334/irsp.133

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806.

Meyer, M. N., & Chabris, C. (2014, July 31). *Psychologists' Food Fight Over Replication of "Important Findings."* Slate Magazine. 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*, *21*(11). https://www.psychologicalscience.org/observer/the-toothbrush-problem

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K. C., Balas, R., Benedict, T., & Corneille, O. (2020). *Incidental attitude formation via the surveillance task: A registered replication report of Olson and Fazio (2001)*. https://doi.org/10/ghwq2z

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., & Antfolk, J. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research Culture and Reproducibility. *Trends in Cognitive Sciences*, *24*(2), 91–93. https://doi.org/10.1016/j.tics.2019.12.002

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

National Academies of Sciences, E. (2019). *Reproducibility and Replicability in Science*. https://doi.org/10.17226/25303

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual*

Review of Psychology, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nosek, B. (2019, June 3). *The Rise of Open Science in Psychology, A Preliminary Report*. https://www.cos.io/blog/rise-open-science-psychology-preliminary-report

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2020a). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., & Errington, T. M. (2020b). The best time to argue about what a replication means? Before you do it. *Nature*, *583*(7817), 518–520. https://doi.org/10.1038/d41586-020-02142-6

Nosek, B. A., & Gilbert, E. A. (2017). *Mischaracterizing replication studies leads to erroneous conclusions* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nt4d3

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*,

7(6), 615–631. https://doi.org/10.1177/1745691612459058

Nosek, Brian. (2019). Strategy for Culture Change. *Strategy for Culture Change*. https://cos.io/blog/strategy-culture-change/

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, *41*, e143. https://doi.org/10.1017/S0140525X18000791

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. https://doi.org/10/f9pdjm

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, M. B., van Assen, M. A., Veldkamp, C. L., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172–182. https://doi.org/10.1037/gpr0000034

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., … Zrubka, M. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268–294. https://doi.org/10.1177/1745691618755704

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*. https://doi.org/10.1037/bul0000294

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science.

Science, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544.

Pawel, S., & Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, *15*(4), e0231416. https://doi.org/10.1371/journal.pone.0231416

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. (2020). *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable*. PsyArXiv. https://doi.org/10.31234/osf.io/n2a9x

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition* (5th edition). Free Press.

Romero, F. (2017). Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, *84*(5), 1031–1043.

Root, P. (2020). *Systematizing Confidence in Open Research and Evidence*. https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10/d5sxt3

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, 1–7.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069. https://doi.org/10/f83jv5

Scheel, A. M., Schijen, M., & Lakens, D. (2020). *An excess of positive results: Comparing the standard Psychology literature with Registered Reports*. https://doi.org/10.31234/osf.io/p6e9c

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schnall, S. (2014). *Clean data: Statistical artifacts wash out replication efforts.*

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology*, *45*(4), 305–306.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001

Sedlmeier, P., & Gigerenzer, G. (1992). *Do studies of statistical power have an effect on the power of studies?*

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell.* Boston: Houghton Mifflin,.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639. https://doi.org/10.1073/pnas.1711786114

Shih, M., & Pittinsky, T. L. (2014). *Reflections on positive stereotypes research and on replications.*

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 1–7. https://doi.org/10.1038/s41562-020-0912-z

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10/gc7qn4

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods*

and Practices in Psychological Science, *1*(1), 115–120.

Soderberg, C. K., Errington, T., Schiavone, S. R., Bottesini, J. G., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2020). *Research Quality of Registered Reports Compared to the Traditional Publishing Model*. MetaArXiv. https://doi.org/10.31222/osf.io/7x9vy

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10/gfx5h3

Spellman, B. A. (2015). A Short (Personal) Future History of Revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (2012). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*. http://www.tandfonline.com/doi/abs/10.1080/00031305.1995.10476125

Sterling, Theodore D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10/gckf9z

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10/b4r4

Tiokhin, L., & Derex, M. (2019). Competition for novelty reduces information sampling in a research game-a registered report. *Royal Society Open Science*, *6*(5), 180934. https://doi.org/10/gf9h3t

TOP Resources—Evidence and Practices. (2016). https://doi.org/None

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2020). *Credibility Beyond Replicability: Improving the Four Validities in Psychological Science*. PsyArXiv. https://doi.org/10.31234/osf.io/bu4d3

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., & Bakos, B. E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317.

Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology: General*, *148*(9), 1628–1639. https://doi.org/10.1037/xge0000663

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A.,

Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, Eric-Jan, Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). *Why psychologists must change the way they analyze their data: The case of psi: comment on Bem (2011).*

Wagenmakers, Eric-Jan, Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wagge, J., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., & Grahe, J. (2018). *A Demonstration of the Collaborative Replication and Education Project: Replication Attempts of the Red-Romance Effect*. PsyArXiv. https://doi.org/10.31234/osf.io/chax8

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). *Intellectual humility: Owning our limitations*. https://doi.org/10/ghwq2x

Wicherts, J. M., 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. https://doi.org/10/g29

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10/bdd4

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, *117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117

Wilson, B. M., & Wixted, J. T. (2018). The Prior Odds of Testing a True Effect in Cognitive and Social Psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197. https://doi.org/10.1177/2515245918767122

Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(20), 10762–10768. https://doi.org/10.1073/pnas.1909046117

Yarkoni, T. (2019). *The Generalizability Crisis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jqw35

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

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018a). Improving social and behavioral science by making replication mainstream: A response to commentaries. *Behavioral and Brain Sciences*, *41*. https://doi.org/10/gjghg5

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018b). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X17001972

Supplementary Information for Nosek et al. (2021)

Author Contribution Statement

B.A.N. Drafted the outline and manuscript sections, collaborated with section leads on conceptualizing and drafting their components, coded replication study outcomes for Figure 1, coded journal policies for Figure 5.

T.E.H. Drafted small sections of the manuscript (in ‘What happens after replication?’ and ‘Evidence of change’). Collected and analyzed data for Figure 5 and Figure S1. Made suggestions and revisions to the manuscript prior to submission.

H.M. Drafted small sections of the supplement (“Are behaviors changing?”) and contributed content to sections of the manuscript (“Evidence of change” section). Compiled, curated, and synthesized data for tables S4, S5, and S6. Contributed to curating data for Figure 1.

A.A. Drafted small sections of the manuscript. Compiled, curated, and analyzed data for Figure 1.

K.S.C. Drafted small sections of the manuscript. Made revisions to the manuscript.

A.D. Drafted small sections of the manuscript. Compiled and analyzed data for Figure 2 and Figure S2.

F.F. Drafted small sections of the manuscript. Made revisions to the manuscript.

J.H. Drafted small sections of the manuscript (“Evidence of Change”) and supplement (“Retractions”). Analyzed data for these sections. Generated Figure 4.

M.K.S. Drafted small sections of the manuscript. Made revisions to the manuscript.

M.B.N. Drafted small sections of the manuscript. Made revisions to the manuscript.

J.R. Drafted small sections of the manuscript. Made revisions to the manuscript.

F.R. Drafted small sections of the manuscript and parts of supplement "What happens after replication?" Made revisions to the manuscript.

A.S. Drafted small sections of the manuscript. Made revisions to the manuscript.

L.S. Drafted small sections of the manuscript (“Evidence of change” section). Contributed to literature review appearing on pages 20-22, and coding appearing in Tables S4, S5, and Figure 4. Made suggestions and revisions to manuscript prior to submission.

F.S. Drafted small sections of the manuscript. Collected and analyzed data for the section on “changes in job advertisements”.

S.V. Drafted small sections of the manuscript. Made revisions to the manuscript.

Summary of Replication Studies

Table S1. Descriptive statistics of Original and Replication Studies summarized in Figure 1. “Multi-site replications” include the series titled “Many Labs” (Ebersole et al., 2016, 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, Skowronski, et al., 2018; O’Donnell et al., 2018; Verschuere et al., 2018; E.-J. 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 Consortium, 2020; Mccarthy et al., 2020; McCarthy, Hartnett, et al., 2018; Moran et al., 2020; Schweinsberg et al., 2016). “Best practice” refers to a prospective replication effort that incorporated preregistration, large samples, and methods transparency for original findings in an effort to increase replicability (Protzko et al., 2020).

	Systematic Replications				
	Soto (2019)	Camerer (2018)	Open Science Collaboration (2015)	Multi-site Replications	“Best Practice” Protzko (2020)
Number of replication outcomes	101	21	94	77	14
Sample size					
Mean (Median) of original studies	1371.82 (468)	72.43 (51)	2526.77 (54.5)	242.84 (82.5)	1608.71 (1544.5)
Mean (Median) of replication studies	1298.08 (1505)	458.19 (243)	5229.11 (70)	4150.40 (3549)	7034.07 (6430.5)
Ratio of Means (Medians)	0.946 (3.22)	6.33 (4.76)	2.07 (1.28)	17.09 (43.02)	4.37 (4.16)
Effect size (r)					
Mean (Median) of original studies	0.296 (0.270)	0.460 (0.388)	0.406 (0.378)	0.335 (0.330)	0.149 (0.132)
Mean (Median) of replication studies	0.239 (0.228)	0.246 (0.149)	0.202 (0.126)	0.176 (0.070)	0.143 (0.133)

Difference of Means (Medians)	-0.061 (-0.045)	-0.241 (-0.253)	-0.222 (-0.264)	-0.169 (-0.266)	-0.006 (0.001)
Statistical Significance					
% of replications significant and in same direction as original	90.1%	66.7%	35.1%	55.8%	100%

Notes: Statistical significance criterion only includes studies for which the original study was statistically significant. Sample includes more than one outcome per study if the replication identified multiple primary outcomes for the evaluation. Sample size refers to the final sample size used in the analysis, i.e., following all exclusions. Ratio of sample size means and medians are replications/original studies. Difference of mean and median effect sizes were calculated by converting *r* to *z*, taking the difference, and then converting *z* to *r*. Negative values indicate that the replication was smaller than the original study. Table S1 includes only replication studies for which we could convert both the original effect and the replication effect to Pearson's r. In addition to the replications reported in Table S1, we identified 48 additional replication studies for which conversion to Pearson's r could not be easily obtained.

Figure 1 and Table S1 include all systematic and multisite replication studies since 2012 that we identified in search of the psychological literature. However, we do not claim that we identified all such studies. Some cells of the table have lower N than the overall count because of missing data or because the reported statistics did not allow effective conversion to a common effect size metric (Pearson's *r*).

There are two known systematic replication studies in neighboring disciplines that were not included in the main text summary:

- (Cova et al., 2018) conducted systematic replications of 37 studies in experimental philosophy and observed statistically significant results in the same direction for 78% with effect sizes 89% of the original study on average.
- (Camerer et al., 2016) conducted systematic replications of 18 studies in experimental economics and observed statistically significant results in the same direction for 61% with effect sizes 66% of the original study on average.

We also excluded (Aczel et al., 2019) from the multisite replications list because of non-comparability of effect sizes between the original and replication studies. We decided to retain (ManyBabies Consortium, 2020) despite there not being a singular original study for testing replication and used a meta-analytic effect size reported from a review of that literature.

What happens after replication?

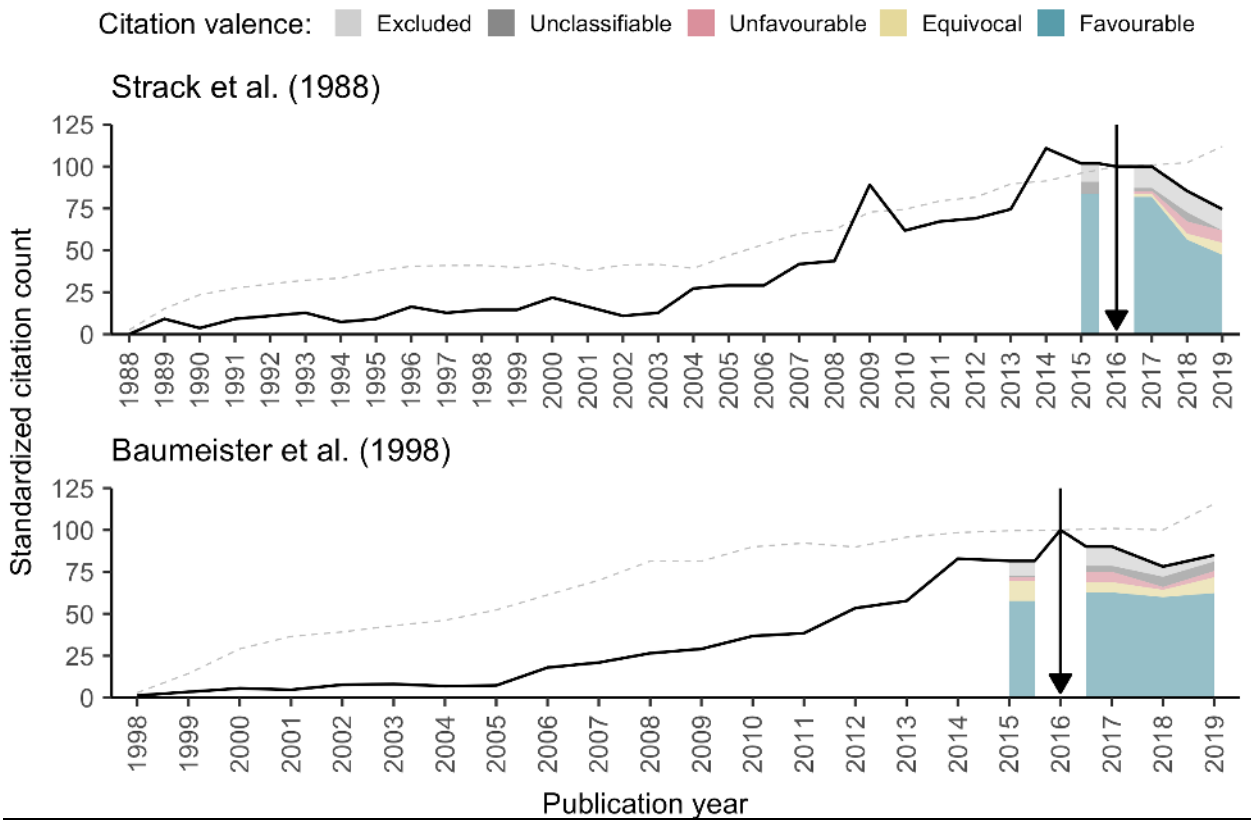
Science is presumed to be self-correcting by progressively refuting falsehoods and improving the validity of scientific knowledge (Laudan, 1981). Replications, robustness tests,

and reproducibility checks support this self-corrective process. Self-correction could begin with individual scientists shifting their judgments about original findings after replication. There is preliminary evidence that psychological scientists would update their judgments if replication evidence is brought to their attention and they are explicitly prompted to reflect on their judgments (McDiarmid et al., 2021). However, whether replications spontaneously prompt scientists to update their judgments in real world settings is less clear.

Research on ego depletion provides a useful case study. Ego depletion refers to the idea that acts of self-control rely on limited resources and using these resources makes subsequent acts of self-control less successful (Baumeister et al., 1998). Hundreds of published studies appeared to support ego depletion with demonstrations of generalizability to different contexts and few replications. Multisite preregistered replications elicited little evidence for ego depletion (Hagger et al., 2016), but the findings were criticized by other ego depletion researchers (Baumeister & Vohs, 2016; Dang, 2016). Meta-analytic evidence also challenged the robustness of ego depletion (Carter et al., 2015; Carter & McCullough, 2014), and was also criticized (Cunningham & Baumeister, 2016) and countered with another meta-analysis (Dang, 2018). Leading researchers organized another preregistered, multisite effort and found little to no support for ego depletion (Vohs et al., 2021).

One indicator of self-correction might be reflected in how studies are cited. An examination of post-replication citation patterns in psychology suggests that even clearly contradictory replication results may only have a modest impact on how original studies are appraised in subsequent academic literature (Figure S1; Hardwicke et al., 2021). In the case of ego depletion, there was a small increase in favorable citations (71% to 73%) to Baumeister and colleagues (1998) and a small increase in unfavourable citations (3% to 5%) from pre-replication (2015) to post-replication (2017-2019).

Figure S1. Standardized annual citation counts (solid line) with citation valence (favourable, equivocal, unfavourable, unclassifiable) illustrated by coloured areas in pre-replication and post-replication assessment periods. Dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Citation valence classifications for the Baumeister et al. case are extrapolated to all articles in the assessment period based on a 40% random sample. Figure by Hardwicke et al. (2021) available at https://osf.io/pxkt2/ under a CC-BY 4.0 license. Data and reproducible analysis code available at https://osf.io/6hsny/ and https://doi.org/10.24433/CO.4445269.v1



In the case of the facial feedback hypothesis by Strack and colleagues (1988), another classic finding with a prominent failure to replicate (E.-J. Wagenmakers et al., 2016), there was a decrease in favourable citations (82% to 71%) and a small increase in unfavourable citations (0% to 6%). These figures suggest modest corrective effects and imply considerable perpetuation of belief in the credibility of the original findings despite contradictory replication results. Even if clearly contradictory replication results do not necessarily undermine the credibility of an original finding (Collins, 1992; Earp & Trafimow, 2015; Maxwell et al., 2015), one might expect relevant counterevidence to be acknowledged and addressed with explicit argumentation. However, Hardwicke et al. observed substantial citation bias: only 27% of post-replication articles citing Strack and colleagues (1988) and 18% of those citing Baumeister and colleagues (1998) also cited the respective large-scale replication. Of those articles that cited the original study favorably and cited the replication, a principled defence of the original study appeared in 50% and 54% of articles respectively. Thus, in these cases, relevant replication evidence has been neglected in post-replication citation patterns.

Part of the explanation for the lack of change in citation patterns may be continuing substantive disagreement about the meaning and relevance of the replication effort and the emergence of new evidence. Working out these boundary conditions and testing alternative hypotheses on the source of a replication failure can lead to productive scientific discovery. For the facial feedback hypothesis, Strack (2016) argued that the Wagenmakers and colleagues (2016) replication attempt was flawed because of the presence of a video camera, which could induce self-awareness and hence suppress emotional responses, among other methodological

concerns. Noah and colleagues (2018) manipulated the presence of a video camera in a preregistered study and found a facial feedback effect in the absence of a video camera which disappeared when the camera was present. However, Wagenmakers and Gronau (2018) criticized that replication evidence as the preregistration underspecified the analyses that support the finding. The authors of the replication disagreed (response by Schul et al. in Wagenmakers and Gronau 2016). Marsh and colleagues (2019) reported the combined data from in-class demonstrations of the facial feedback hypothesis to provide a large sample test and observed a small-to-medium effect size, in line with the original report. In their setting, participants were made aware of both experimental conditions and had many observers present. This seems potentially problematic for testing the hypothesis given earlier concerns about demand characteristics and induced self-awareness, but not all researchers agree. Strack suggested that this approach does test the hypothesis (Strack, personal communication, February 14, 2021). Optimistically, the viability of group administration increases the feasibility of conducting large sample, high-powered tests of the hypothesis in the future.

Ideally, replication studies would produce constructive responses, including updating beliefs among the scientific community and empirically testing possible explanations for inconsistent results. Indeed, critical self-reflection is a hallmark of what makes scientific communities trustworthy (Vazire & Holcombe, 2020). More metascientific research is needed to understand under what conditions scientific communities engage in productive or counterproductive responses, and how fields respond to replications of phenomena over longer timespans. The recent reform movement in psychology has provided examples of a wide range of responses and some hints to the features of productive debates. When scientific communities are responding productively--even with strong or potentially unresolvable differences of opinion among some contributors--we observe some of the following consequences for the discipline:

(1) Critical reevaluation of theories, hypotheses, data, and auxiliary assumptions (e.g., about methods, designs, or statistical practices)

(2) Open, collaborative, and constructive dialog in debates where even researchers holding strongly disagreeing positions are able to maintain assumptions of best intent and avoid harmful retaliation or ad-hominem attacks

(2) Refined estimates of effect sizes and their practical implications (e.g., Byers-Heinlein et al., 2020)

(3) Systematic and rigorous exploration of potential boundary conditions and moderator variables;

(4) Investigations to evaluate findings theoretically related to the original findings

(5) Redirection of research efforts to more promising avenues

Recovering findings following a failure to replicate

Debates about the facial feedback hypothesis and ego depletion are related to a topic briefly reviewed in the main text about whether problems in replication studies are occasionally or frequently the explanation for failures to replicate. There we reported evidence from Ebersole and colleagues (2020) that 10 findings replicated in Open Science Collaboration (2015) did not show evidence of greater replicability following expert peer review of the experimental protocols

prior to conducting the study. Two other studies provide relevant evidence about whether incompetence or other weaknesses due to replication team efforts are responsible for failures to replicate.

First, Klein and colleagues (2019)investigated the role of expertise in conducting a replication of a key finding from Terror Management Theory (TMT; Greenberg et al., 1994). Klein and colleagues randomly assigned labs with variable experience with TMT to one of two conditions: [1] review the original paper and generate a replication protocol themselves or [2] use an experimental protocol designed with the input of TMT experts. Across a total of 21 labs and 2,220 participants, there was little evidence for the original finding in either condition. The average effect size across conditions ranged from Hedge's *g* of .03 to .07 depending on the exclusion criteria applied, and the 95% confidence intervals overlapped 0. Moreover, whether the protocol was self-designed by the lab or expert-designed did not moderate the size of the observed result (*p*'s = .75, .61, .73). The observed effect size was considerably smaller than the original finding, Cohen's *d* = 1.34 (Greenberg et al., 1994).

Second, Ebersole and colleagues (2016) failed to replicate a classic finding related to the elaboration likelihood model (ELM; *N* = 114, original effect size f^2 = 0.20, 95%CI [0.06, 0.41]; Cacioppo et al., 1983; Replication N = 2365, f^2 < 0.001, 95%CI [0, 0.002]). Petty and Cacioppo (2016) suggested hypotheses for why the finding might have failed to replicate including possible improvements to the experimental design. Luttrell and colleagues (2017) implemented presumed improvements and found suggestive evidence (*p* = 0.03) for an interaction between the Ebersole et al. protocol (*N* = 106, f^2 = 0.001, 95%CI [0, 0.057]) and their revised protocol (*N* = 108, f^2 = 0.07, 95%CI [0.003, 0.196]). However, even the revised protocol effect size was considerably smaller than the original finding which was outside of even the wide confidence interval. Ebersole and colleagues (2017) replicated the Luttrell et al. design with a very large sample (*N* = 1219) using many labs and failed to replicate the interaction supporting the conclusion that the revised protocol improved replicability (*p* = 0.135), but did find some support for a positive effect though at 1/8th the original effect size (f^2 = 0.025, 95%CI [0.006, 0.056]). This scholarly examination of replicability provides the most promising possibility that a finding might be recovered following expert revisions to a failure to replicate. In strictly null hypothesis testing terms, the original study provided support for the finding, the replication did not, the expert-revised methodology did, and then the replicator team's use of the expert-revised methodology did. This suggests the value of expertise in improving replicability in this context. On the other hand, the lack of statistical significance comparing conditions fails to support that conclusion and the dramatic decline in effect size compared to the original study is even more problematic for the notion that replicability is a function of expert participation. Nevertheless, it is always a possibility that failures in implementation of replications could account for null or reduced effects, just as it is always a possibility that failures in implementation of original research could account for positive or inflated effects (Greenwald, 1975). As such, it is not possible to draw a conclusion about the role expertise and implementation quality for replication outcomes in general.

Prediction surveys, betting markets, elicitation techniques, and machine learning

Table S2. Descriptive statistics for prediction surveys, betting markets, structured elicitations, and machine learning outcomes summarized in Figure 3.

	Prediction markets	Surveys	Structured elicitations
Number of replication outcomes	123	123	25 (Wintle et al., 2021)
Replication projects predicted	Open Science Collaboration (2015) (41 studies), Camerer et al. (2016) (18 studies), Camerer et al. (2018) (21 studies), Klein et al. (2018) (24 studies), Ebersole et al. (2020) (20 studies)	Open Science Collaboration (2015) (41 studies), Camerer et al. (2016) (18 studies), Camerer et al. (2018) (21 studies), Klein et al. (2018) (24 studies), Ebersole et al. (2020) (20 studies)	Klein et al (2014) (1 study), Open Science Collaboration (2015) (9 studies), Ebersole et al. (2016) (1 study) Camerer et al. (2018) (6 studies), Klein et al. (2018) (8 studies)
Number of participants	31-114	31-114	25 (5 groups of 5 participants each)
Classification Accuracy	72%	64%	84%
Mean of predictions	57.2	56.8	51.7
Median of predictions	60.8	56.1	54.5
Range of predictions	6.53-95.5	21.7-88.7	19.2-75.8
Standard deviation of predictions	23.2	16.9	17.3

Notes: Classification accuracy is calculated as the share of correctly predicted replication outcomes based on dichotomizing prices or elicited probabilities above 50 as predicting successful replication. The table shows the unweighted mean across scoring strategies; Structured elicitations had 16 preregistered aggregation methods with performance range of 80% to 88%.

Prediction markets are information aggregation tools (Plott & Sunder, 1988), where participants typically trade contracts with clearly defined outcomes. With some caveats

(Fountain & Harrison, 2011; Manski, 2006) prices for contracts with binary events can be interpreted as the probabilities that the market assign the events. Prediction markets have been used to assess replication outcomes starting with Dreber et al. (2015), and there are now a handful of papers on this topic. In these markets, participants typically bet on whether a study or hypothesis will replicate or not (binary event), where successful replication is most often defined as a result in the same direction as the original study that is statistically significant ($p<0.05$ in a two-sided test). A contract is worth $1 if the study replicates, and $0 if it does not replicate. In these markets, the price of a study/hypothesis is interpreted as the probability that the market assigns the replication outcome to be successful, not that the hypothesis is true. If the contract is priced for example 50 cents, this means that a participant should buy the contract if they think that the study has more than 50% probability of replicating, whereas if they think the study has less than a 50% probability of replicating, they should short sell the contract. In the simplest type of analysis, market prices above 50 can be interpreted to mean that the market thinks that the study will replicate, and prices below 50 that the study will not replicate. There are several reasons for why these markets could work well in aggregating information and performing well in predicting replication outcomes, including that predictions are incentive compatible and that participants based on prices can learn something about other people's beliefs about the study replicating, and thus update their beliefs. In these projects, the markets are open for some time period (10-14 days) and participants are endowed with USD 50-100.

Dreber et al. (2015) used prediction markets to predict replication outcomes for 41 (actually 44 but 3 replications were not finished) of the RPP studies. Participants were invited through the Open Science Framework and the RPP network. In the instances where participants were also replicators, they were not allowed to bet on their replications. The mean prediction market price was 55%, with a range of 13-88%, indicating that about half of the studies were expected to replicate. For the 41 studies explored, 16 (39%) replicated and 25 (61%) did not replicate. Using the price of 50 cutoff, the markets correctly predicted 29 of 41 replications (71%). Camerer et al. (2016) replicated 18 studies in experimental economics and found that 11 studies replicate. They also included prediction markets, with mainly experimental economists participating, and found that the market predicted that all studies would replicate in the sense that all prices were above 50, with the mean market price being 75%. Camerer et al. (2018) also had prediction markets for the 21 studies published in Nature and Science that they replicated. Participants were mainly researchers in psychology and economics and related fields. In this study 13 studies replicated, and the markets correctly predicted that these 13 would replicate but also that three failed replications that would replicate. Forsell et al (2018) also included prediction markets for 24 (28 but some of the replications changed) replicated in the Many Labs 2 project. Unlike in the other projects, replication success was here defined as an effect in the same direction that is statistically significant at $p<0.001$. Prediction markets, mainly including psychologists, correctly predicted 18 of 24 (75%) replication outcomes.

Gordon et al., (2021) pooled the data from Dreber et al. (2015), Camerer et al. (2016, 2018) and Forsell et al. (2018) and found that for the 103 replication outcomes for which they have prediction market data and survey beliefs, prediction markets correctly predict 75 outcomes (73%) when interpreting market prices above 50 as the market believing the study will replicate.

In addition, Ebersole et al. (2020) added prediction markets to the Many Labs 5 project, where researchers were asked to predict the replication outcomes of the 10 RP:P protocols and the 10 revised protocols. While two revised protocols replicated the original effect, the markets (consisting mainly of psychologists) believed that all but two other versions (RP:P and revised protocol) for another study would not replicate. Neither the prediction markets nor the survey performed well in predicting replication outcomes in this study.

In Forsell et al. (2018) participants were also invited to predict relative effect sizes of the replications compared to the original studies. These markets attracted little trading - participants could choose whether to invest their endowments mainly in the binary markets or the effect size markets, and most chose the former.

In the figures we show here we have combined the pooled results from Gordon et al. (2021) and the results from Ebersole et al. (2020). This leads to a sample of 123 prediction-replication pairs for which there are both survey and market predictions. With the dichotomous criterion of prices above 50 anticipating replication success, the prediction markets successfully predict 88 of 123 (72%) while surveys correctly predict 79 of 123 (64%) results.

There is also an example of lay judgments performing well in predicting a subset of these replication outcomes (Hoogeveen et al. 2020) - when 233 participants without a PhD in psychology were given the hypotheses for 27 of the replications in Camerer et al. (2018) and Klein et al. (2018), they predicted the outcomes better than chance (59%). When given additional information, their prediction performance reached 67%.

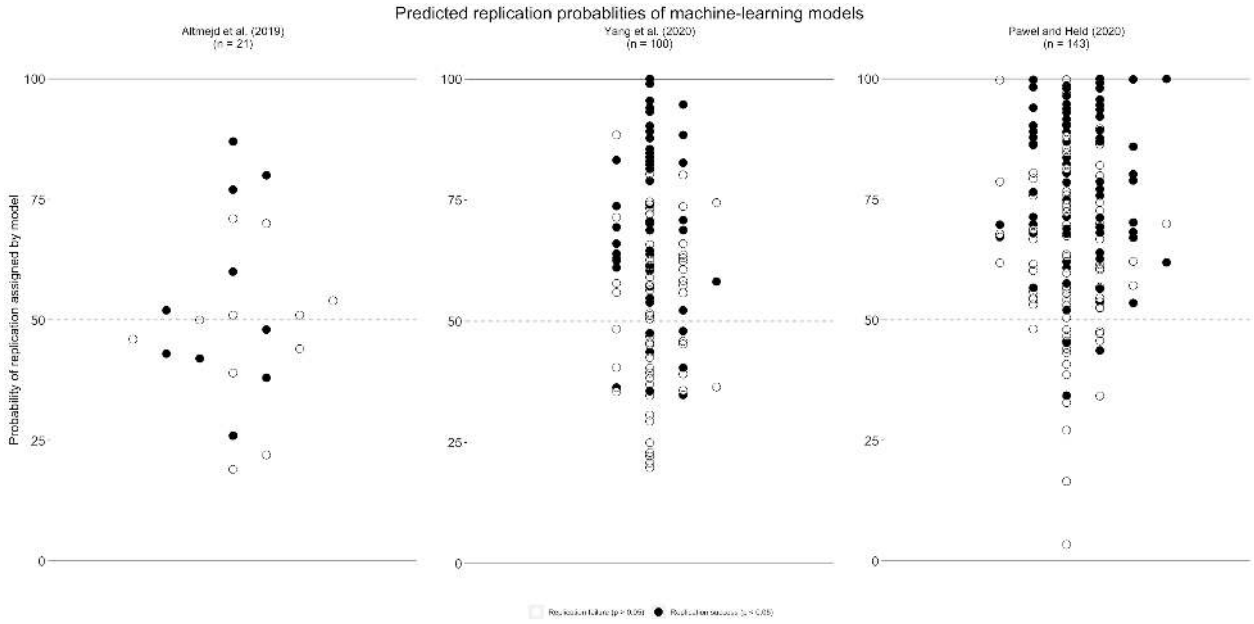
In addition, there are several recent studies using similar prediction surveys with and without monetary incentives to predict different types of replication outcomes, e.g. Landy et al. (2020) where the focus is on conceptual replications and Tierney et al., (2020) who propose a “creative destruction” approach to replication where theories or frameworks are pitted against each other in the replications. These studies, with mainly researchers predicting replication outcomes, again suggest some albeit imperfect “wisdom of crowds”.

Structured elicitation methods report high prediction accuracy, though the number of studies they have been benchmarked against is currently low. Using the IDEA protocol, Wintle et al (2021) asked participants to predict the replicability of 25 replication studies selected from five previously completed replication projects Open Science Collaboration (2015), Klein et al. (2014, 2018), Camerer et al. (2016) and Ebersole et al. (2016). They first excluded widely publicised studies, and randomly selected from the remainder of studies in those five projects. The elicitation was conducted face-to-face, and participants could not ‘look up’ the outcomes. Classification accuracy (the ability to correctly discriminate binary classifications replication outcomes as successes or failures) was 84%, based on an unweighted mean of all second round judgements. (Other preregistered aggregation models had classification accuracies between 80-88%.) On average it took the groups of 5 participants 28 minutes to complete an evaluation of each original study.

When it comes to predictions from machine learning methods (Figure S2), Altmejd et al. (2019) use a black-box statistical approach to develop and train predictive models to predict the outcomes of Open Science Collaboration (2015), Klein et al. (2014), Camerer et al. (2016) and Ebersole et al. (2016). They look at 131 original study-replication pairs, for which they have prediction market prices for 55 pairs which are used as a benchmark for evaluating the model. They also do a pre-registered out of sample test of the 21 study-replication pairs in Camerer et

al. (2018). The authors include independent variables such as objective characteristics of the original studies (standardized effect size and p-value), as well as contextual information such as highest seniority and gender composition of the replication team, length of the paper and citations. The accuracy levels of the predictive models are similar to the prediction markets for the binary replication indicator (whether the replication result is in the same direction with $p<0.05$ or not). Altmejd et al. (2019) also try to predict relative effect sizes, and the predictive models for these do not perform as well as for the binary replication indicator. The results suggest that statistical properties like 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.

Figure S2. Predictions of replication outcomes across three machine learning projects.



Yang et al. (2020) also predict replicability with machine learning models, starting with the Open Science Collaboration (2015) as training data, and doing out-of-sample tests on a more extensive set of replications from psychology as well as economics compared to Altmejd et al. They compare predictive models trained on either the original papers' narrative (text), reported statistics or both narrative and reported statistics. The model's accuracy is higher when trained on narrative than when trained on reported statistics, and the results also suggest that higher word combinations (ngrams) correlate with replication. As in Altmejd et al. (2019), the model performs as good but not better than the prediction markets for the sample of 100 studies for which they do this comparison.

Pawel and Held (2020) use a different type of forecasting approach using the original studies' information and the replication studies' sample size only for (subsets of) the data in Open Science Collaboration (2015), Camerer et al. (2016), Camerer et al. (2018) and Cova et al. (2018). For the studies for which there are prediction market forecasts, the forecasts from Pawel and Held's four different statistical methods perform as well as or worse than the prediction markets in predicting replication outcomes.

Are behaviors changing?

Sample Size

Conducting studies with high statistical power supports replicability. Statistical power and sample sizes, which proxy statistical power, have modestly increased in psychology over the last sixty years. Multiple assessments of statistical power in psychology suggest that statistical power has been consistently low and has not substantially increased from 1955 to 2014 (Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1992; Smaldino & McElreath, 2016; Stanley et al., 2018). Meta-researchers have characterized the median sample sizes of studies sampled from psychology journals from 1977 to 2018 (Fraley & Vazire, 2014; Kossmeier et al., 2019; Marszalek et al., 2011; Reardon et al., 2019; Schweizer & Furley, 2016). Our synthesis of these data (available at osf.io/nb2fv) suggests that sample sizes in almost all research areas have been largely stable. Consistent with this trend, a survey of research published in top psychology journals of confidence intervals - which are negatively related to sample size - did not get smaller from 2002 to 2013 (Brand & Bradley, 2016). However, sample size is merely a proxy for power. If people are using simpler designs with similar Ns (as suggested by Kossmeier et al., 2019; Schweizer & Furley, 2016), then power may be increasing despite the stability in sample size.

There is not clear evidence that longitudinal patterns of statistical power and sample sizes over time differ as a function of research area. Sample sizes do differ on average as a function of research contexts; sample sizes tend to be relatively lower in highly resource-intensive research areas (e.g., neuroscience, Button et al., 2013), in experimental designs relative to correlational designs (Kossmeier et al., 2019; Schweizer & Furley, 2016), and, likely, in studies that aren't and can't be conducted online (Anderson et al., 2017). Apparent differences in longitudinal patterns, for example, in personality psychology relative to sports and exercise psychology, may be due to differences in the complexity and design of personality research rather than to a collective change in research practices among personality psychologists.

Retractions

Retraction is a mechanism for removing flawed or fabricated results from the research literature. We used the Retraction Watch Database (http://retractiondatabase.org/) to explore trends in retractions in psychology (analysis code is available at osf.io/w6fev).

The Retraction Watch database contains records of more than 15,000 retractions of scientific journal articles and conference papers from 1927 to today. Due to variation in retraction notices, there are a few inconsistencies in the database. The date of a retraction is often the date that the article's retraction notice was published, but is sometimes the original publication date. The subject area categorization does not always align with our conceptualization of psychological science as inclusive of behavioral psychology, clinical psychology, and neuroscience. Consequently, our analyses - which use the "(SOC) Psychology" database tag - may undercount retractions in the psychological sciences (e.g., a retraction in the journal Psychological Science related to hand sanitizer use in the workplace is tagged as "(HSC) Public Health and Safety"). Before conducting the analyses below, we

removed two mega-retraction events from the database: 1) the simultaneous retraction of 7,500 articles and conference abstracts originally published by the Institute of Electrical and Electronics Engineering and 2) the simultaneous retractions of 434 conference abstracts originally published by the Journal of Fundamental and Applied Sciences. Prior to their removal, these two events accounted for 36% of all retractions with the tag “(SOC) Psychology”.

We assessed 14,966 retractions in psychology (i.e., retractions tagged ”(SOC) Psychology”). The number of retractions in psychology is more than other social sciences (communication, education, political science, or sociology). In the years 2002-2010, the mean number of retractions per year was 3.4 (SD = 3.1), and the highest number of retractions in a single year was 11 in 2009. In 2011, the fraud of Diederik Stapel was discovered; since that time, psychology journals have retracted an average of 35.7 articles per year (SD = 11.8).

Outside of major investigations of specific researchers, retraction for falsification or fabrication are relatively rare in psychology; more common are retractions citing concerns about the reliability of data and results. The Retraction Watch database lists the non-exclusive reasons for retraction of each article as described in the article’s retraction notice. 14.2% have no listed reason. The most commonly cited reason for retraction was errors in analyses, data, methods, results, or conclusions (21.1%). Other common reasons were concerns regarding the accuracy or validity of the data or results (20.6%), falsification or fabrication (20.4%), plagiarism or other issues regarding referencing (14.8%), and duplication of articles (i.e., “self-plagiarism,” 12.2%).

The median lag time between publication and retraction in psychology is 1.9 years. The average lag time is considerably longer, 3.6 years. The slowest retractions have just appeared in 2020, retracting the work of Hans Eysenck and Ronald Grossarth-Matticek after more than twenty years.

Researchers have argued that the increased rate of retractions in psychology over time is caused by improved journal oversight rather than increased incidence of fraud, citing more journals issuing retractions but not more retractions per journal (Brainard, 2018). The increase of retractions over time in psychology are likely not due to increasing misconduct, but instead due to an increasing political will and ability to detect research misconduct and impossible statistical results using tools like GRIM and SPRITE (Brown & Heathers, 2017; Heathers et al., 2018).

Open science practices

Table S3. Proportion of surveyed psychology researchers who engaged in open science practices from 2016 to 2019

Source	Year	Targeted population of psychology researchers	N	Open data	Open materials	Preregistration
(CWTS, 2017)	2016	All	59	0.55		

(Washburn et al., 2018)	2017	Social and Personality	1035	0.56		0.27
(Houtkoop et al., 2018)	2017	All	780	0.40		
(Christensen et al., 2019)	2017	All	86	0.60*	0.44	0.51
(Beaudry et al., 2019)	2019	Australian researchers	45	0.27	0.20	0.36
(Makel et al., 2019)	2019	Education	1488	0.46	0.59	0.54
(Van den Akker et al., 2020)	2019	Emotion researchers	144			0.57
Min				0.27	0.20	0.27
Median				0.50	0.44	0.51
Max				0.60	0.59	0.57
Weighted mean				0.48	0.57	0.44

Note. *Christensen et al. (2019) report the proportion of psychologists who reported sharing data or analysis code. Year is the effective year of the estimate, which is the year of the survey end date if reported or otherwise the year of the latest possible date based on available information (e.g., initial preprint date, journal submission date). For all open science behaviors, the question stems in most studies asked researchers whether they had ever performed the behavior and one asked in reference to a specific project (CWTS, 2017). CWTS stands for Leiden University's Centre for Science and Technology Studies. The reported estimates for Houtkoop et al. (2018) were extracted from a figure. Empty cells indicate that the survey did not assess that behavior. Table data are from osf.io/jsu4r. Additional documentation is available at osf.io/pqv73. Code to produce the table is at osf.io/3dkux.

Table S4. Proportion of audited psychology research using open science practices from 2013 to 2018

Source	Year	Targeted psychology research	N articles	Open data	Open materials	Preregistration
(Vanpaemel et al., 2015)	2012	General	394	0.38*		
(Hardwicke et al., 2020)	2014-2017	General	250	0.02	0.14	0.03
(Obels et al., 2020)	2014-2018	Registered Reports	62	0.65	0.60	
(Vassar et al., 2020)	2013-2018	Clinical	394	0.00		
Min				0.00	0.14	0.03
Median				0.20	0.37	0.03
Max				0.65	0.60	0.03
Weighted mean				0.18	0.23	0.03

Note.Year is the effective year of the estimate, which is the year or range of years that the sampled literature was published. *Vanpaemel et al. (2015) report the number of articles for which data was provided following a request. Year is the effective year of the estimate, which is the most recent year of the audit study range. Empty cells indicate that the audit did not assess that behavior. Table data are from osf.io/jsu4r. Additional documentation is available at osf.io/pqv73. Code to produce the table is at osf.io/3dkux.

Questionable Research Practices

Table S5. Proportion of surveyed psychology researchers who had engaged in QRP behaviors from 2010 to 2019

First author	Year	Targeted population of psychology researchers (subsample)	N	At least one	Dropped DVs	Continued data collection	Dropped conditions	Stopped data collection	Rounded p-values	Dropped studies	Excluded data	HARKed	Claiming generalization	Falsified data
Bosco	2010	OB	53									0.38		
John	2011	American researchers (control)	466	0.91	0.63	0.56	0.28	0.16	0.22	0.46	0.38	0.27	0.03	0.01
John	2011	American researchers (experimental truth instruction)	970	0.94	0.66	0.58	0.27	0.22	0.23	0.50	0.43	0.35	0.04	0.02
Fiedler	2015	German	1138		0.33	0.32	0.25	0.05	0.22	0.42	0.40	0.46	0.03	0.03
Banks	2015	Management (published in high-impact journal)	405						0.12	0.55	0.27	0.51		0.00
Banks	2015	Management (published in low-impact journal)	318						0.10	0.44	0.30	0.48		0.00

Krishna	2016	German psychology students	207	0.56	0.06	0.02	0.08	0.02	0.10	0.28	0.16	0.10	0.03	0.03
Agnoli	2016	Italian researchers	219	0.88	0.48	0.53	0.16	0.10	0.22	0.40	0.40	0.37	0.03	0.02
Héroux	2016	Electrical brain stimulation researchers	154		0.14		0.13				0.09			
Motyl	2017	Social and Personality	1166		0.63	0.56	0.28	0.16	0.22	0.46	0.38	0.27	0.03	0.01
Fox	2017	American researchers	303	0.18										
Wolff	2018	Ego depletion researchers	277		0.21		0.03				0.17			
Janke	2018	German PhD students and postdocs	217	0.86	0.56	0.23	0.24	0.04	0.08	0.43	0.42	0.41	0.08	
Makel	2019	Education (quantitative only)	1218			0.29			0.29	0.62	0.25			
Makel	2019	Education (qualitative and quantitative)	1488									0.46		
Swift	2019	Clinical (faculty)	164	0.65	0.12	0.11	0.03	0.04	0.13	0.35	0.12	0.14	0.01	0.00

Swift	2019	Clinical (PhD students)	110	0.48	0.04	0.05	0.03	0.02	0.08	0.26	0.09	0.04	0.01	0.01
Rabelo	2019	Brazilian researchers	232	0.88	0.22	0.22	0.34	0.10	0.18	0.55	0.20	0.09	0.04	0.01
Min				0.18	0.04	0.02	0.03	0.02	0.08	0.26	0.09	0.04	0.01	0.00
Median				0.86	0.28	0.29	0.20	0.08	0.18	0.44	0.27	0.36	0.03	0.01
Max				0.94	0.66	0.58	0.34	0.22	0.29	0.62	0.43	0.51	0.08	0.03
Weighted mean				0.78	0.46	0.40	0.23	0.12	0.21	0.48	0.32	0.36	0.03	0.02

Notes: Year is the effective year of the estimate, which is the year of the survey end date if reported or otherwise the year of the latest possible date based on available information (e.g., initial preprint date, journal submission date). The exact descriptions of behaviors follow (John et al., 2012) closely, but differ somewhat in Fiedler & Schwarz, 2016; Héroux et al., 2017; Swift et al., 2020; Wolff et al., 2018. The question stems in most studies asked researchers whether they had ever engaged in the focal behavior, fewer studies asked about frequency (Janke et al., 2019; Makel et al., 2019; Motyl et al., 2017), specific projects (Bosco et al., 2016; Krishna & Peter, 2018; Wolff et al., 2018) or over the past year (Fox et al., 2018). The reported estimates for Fielder and Schwarz (2016) were extracted from a figure. Empty cells indicate that the survey did not assess that behavior. Total sample size is 7,887 after accounting for overlapping samples in the two Makel 2019 rows to avoid double counting. Table data are from osf.io/jsu4r. Additional documentation is available at osf.io/pqv73. Code to produce the table is at osf.io/3dkux.

Table S6. Description of QRP behaviors in Table S5

QRP behavior	Description, with substantive variations in wording across studies noted
At least one	For studies that asked psychologists about whether they had ever engaged in the ten behaviors from John et al. (2012), the proportion that reported engaging in at least one.
Dropped DVs	In a paper, failing to report all of a study's dependent variables. In Fielder and Schwarz (2016) and Swift (2020), asked about failing to report those that are relevant for a finding.
Continued data collection	Deciding whether to collect more data after looking to see whether the results were significant. In Fielder and Schwarz (2016) and Swift (2020) deciding whether to collect more data in order to render non-significant results significant.
Dropped conditions	In a paper, failing to report all of a study's conditions.
Stopped data collection	Stopping collecting data earlier than planned because one found the result that one had been looking for. In Fielder and Schwarz (2016) and Swift (2020): stopping collecting data earlier than planned because the expected results concerning a specific finding were already obtained.
Rounded p-values	In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .05 is less than .05). In Makel et al. (2019): rounding off a *p* value or other quantity to meet a pre-specified threshold (e.g., reporting *p*=.054 as *p*=.05 or *p*=.013 as *p*=.01)
Dropped studies	In a paper, selectively reporting studies that "worked." In Banks et al. (2015): selectively reporting hypotheses that "worked." In Makel et al. (2019): not reporting studies or variables that failed to reach statistical significance (e.g.,$p \leq .05$) or some other desired statistical threshold
Excluded data	Deciding whether to exclude data after looking at the impact of doing so on the results. In Makel et al (2019): deciding to exclude data points after first checking the impact on statistical significance (e.g., $p \leq .05$) or some other desired statistical threshold desired statistical threshold
HARKed	In a paper, reporting an unexpected finding as having been predicted from the start. In Makel et al. (2019): reporting an unexpected finding or a result from exploratory analysis (i.e., not explicitly planned in advance) as having been predicted from the start.

Claiming generalization	In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do). Fielder and Schwarz (2016) asked about claiming that results are unaffected by demographic variables (e.g., gender) although one is actually unsure (or knows that they do).
Falsified data	Falsifying data

Policy changes by journals

One approach to address concerns about research credibility is to leverage academic journal policies to promote potentially beneficial research activities, such as research transparency, whilst discouraging potentially deleterious research activities, such as selective reporting. We conducted an empirical investigation in order to collate, summarise, and describe the policies of a selection of psychology journals in relation to ten pre-defined standards (Table S7) related to research transparency and reproducibility. Eight of the standards - known as the Transparency and Openness Promotion (TOP) guidelines - were originally devised in 2014 by the Center for Open Science (COS) in conjunction with a group of experts, largely from the social and behavioral sciences, including journal editors and funding agency representatives (Nosek et al., 2015) and are maintained by COS and a committee of stakeholders (http://cos.io/top/). Two additional standards ("Publication Bias" and "Open Science Badges") were added when COS introduced the "TOP Factor", a metric to quantify journal adherence to open science principles (https://perma.cc/7LSM-PHVL). Although the standards cover some core aspects of transparency and reproducibility, they are not comprehensive. For example, they do not address policies related to statistical power, conflict of interest statements, or some novel initiatives like the 'pottery barn replication rule' in which journals commit to publishing replications of studies previously published in their archives (https://perma.cc/SP2P-7Y5Q). The main reason for adopting the existing TOP standards as outcome variables in the present study (as opposed to developing new standards or incorporating additional standards) is efficiency - COS is already in the process of extracting and collating journal policies related to these standards, so we can (a) partly capitalize on their existing database; and (b) contribute additional information to that database.

We assessed the prevalence of the ten TOP standards across the field of psychology and the frequency of adoption by some highly influential psychology journals. To address the former goal, we examined a sample of 40 journals randomly selected from amongst all journals belonging to the 'psychology' category defined by Web of Science (Clarivate Analytics). To address the latter goal, we examined the top-5 journals according to their 2019 Journal Impact Factor (JIF) in each of ten subfields of psychology (i.e., 50 journals). Whilst the use of any single quantitative metric as a proxy for journal influence has limitations, we only intended to use JIF as a practical selection criterion that captured a set of highly influential journals. In total with both samples combined we examined 90 journals. Data collection began in September 2020, and was completed by December 30th, 2020.

Table S7. TOP Factor rubric (obtained from https://perma.cc/CT2T-5G3F?type=image on August 25, 2020).

	Policy level			
Policy	**0**	**1**	**2**	**3**

Data citation	No mention of data citation.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article requires appropriate citation for data and materials used consistent with the journal's author guidelines.	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data transparency	Data sharing is encouraged, or not mentioned.	Articles must state whether or not data are available. Requiring a data availability statement satisfies this level.	Articles must have publicly available data, or an explanation why ethical or legal constraints prevent it.	Articles must have publicly available data and must be used to computationally reproduce or confirm results prior to publication.
Analytical code transparency	Code sharing is encouraged, or not mentioned.	Articles must state whether or not code is available. Requiring a code availability statement satisfies this level.	Articles must have publicly available code, or an explanation why ethical or legal constraints prevent it.	Articles must have publicly available code and must be used to computationally reproduce or confirm results prior to publication.
Materials transparency	Materials sharing is encouraged, or not mentioned.	Articles must state whether or not materials are available. Requiring a materials availability statement satisfies this level.	Articles must have publicly available materials, or an explanation why ethical or legal constraints prevent it.	Articles must have publicly available materials and must be used to computationally reproduce or confirm results prior to publication.
Reporting guidelines	No mention of reporting guidelines.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Study prereg	Journal says nothing.	Articles will state if work was preregistered.	Article states whether work was preregistered and, if so, journal verifies adherence to preregistered plan.	Journal requires that confirmatory or inferential research must be preregistered.

Analysis prereg	Journal says nothing.	Articles will state if work was preregistered with an analysis plan.	Article states whether work was preregistered with an analysis plan and, if so, journal verifies adherence to preregistered plan.	Journal requires that confirmatory or inferential research must be preregistered with an analysis plan.
Replication	Journal says nothing.	Journal encourages submission of replication studies.	Journal will review replication studies blinded to results.	Registered Reports for replications as a regular submission option.
Publication Bias	Journal says nothing.	Journal states that significance or novelty are not a criteria for publication decisions.	Journal will review studies blinded to results.	Journal accepts Registered Reports for novel studies as a regular submission option.
Open science badges	Journal says nothing.	Journal awards 1 or 2 open science badges.	Journal awards all 3 open science badges.	*

*There is no information in this cell because TOP does not specify a level 3 for the open badges policy.

Method

Design

Retrospective observational study with a cross-sectional design. Outcome variables are shown in Table S7.

Obtaining extant data and bibliographic information

On August 13th, 2020, we downloaded the databases described in Table S8. Note that Web of Science classifies journals into subject categories, ten of which pertain to the field of psychology (https://perma.cc/9V5T-PH2U). A journal can belong to more than one subject category.

Existing data in TOP-d pertaining to the journals in the sample was extracted. An R script documenting this procedure (https://osf.io/kdnwg/) and the data (https://osf.io/3fn6e/) are available.

Table S8. Assorted online databases used to obtain the samples of journals. The copies we downloaded are available from the links in the "access link" column.

Database	**Description**	**Acronym**	**Source**	**Date last updated[#]**	**Access link**
TOP Factor database	Extant data extracted for assessment of journal TOP Factor	TOP-d	Open Science Framework https://osf.io/qatkz/	August 11, 2020	https://osf.io/5sbg7/
Web of Science Social Sciences Citation Index master list (psychology journals only)	List of all journals included in the Web of Science Social Sciences Citation filtered to remove all entries that did not contain the term 'psychology' in the subject category field	WOS-psych-d	Clarivate Analytics https://mjl.clarivate.com/collection-list-downloads	July 21, 2020	https://osf.io/6wsj3/
Clarivate Analytics Journal Citation Reports	List of journals ranked by 2019 Journal Impact Factor for each of the ten Web of Science subject categories pertaining to the field of psychology*	JCR-psych-d	Clarivate Analytics Journal Citation Reports https://perma.cc/T2CT-3QE6	Jun 29, 2020	https://osf.io/vde48/

[#]According to the database website.

*These categories are "Psychology, Applied", "Psychology, Biological", "Psychology, Clinical", "Psychology, Developmental", "Psychology, Educational", "Psychology, Experimental", "Psychology, Mathematical", "Psychology, Multidisciplinary", "Psychology, Psychoanalysis", and "Psychology, Social".

Sample

Definition of samples

There were two samples of journals:

1. *Random sample*: 40 journals randomly selected from amongst all psychology journals indexed in the Web of Science Social Sciences Citation Index (WOS-psych-d).

2. *High-impact sample*: The top-5 journals according to 2019 Journal Impact Factors in each of ten subfields of psychology defined by Web of Science (JCR-psych-d; i.e., 50 journals).

Analysis scripts documenting the sampling and screening process are available (https://osf.io/4fqzj/).

Sample size justification

The purpose of the high impact sample was to enable us to gauge the frequency of the outcome variables at some of the most influential psychology journals. To achieve this, we selected the top-5 journals in each subject category of psychology as we intuitively believed this to be informative and within our workload capacity.

The purpose of the random sample was to enable us to estimate the prevalence of the outcome variables across the field of psychology. A precision analysis performed to inform our decision about target sample size is available in the pre-registered protocol (https://osf.io/n9325/).

Inclusion and exclusion criteria

- We only included journals classified as English language in WOS-psych-d because we did not have the resources to achieve high-quality translation. 71 non-English language journals were excluded from WOS-psych-d prior to sampling.
- We did not include journals that do not typically publish empirical research involving primary data (from herein 'non-empirical journals') because the outcome variables are less applicable to these journals.
    - To exclude non-empirical journals, any journals identified as non-empirical during the sampling process were replaced.
        - In the case of the random sample, replacement involved randomly drawing another journal from amongst remaining journals in the WOS-psych-d. Three such replacements were required.
        - In the case of the high-impact sample, replacement involved selecting the next ranked journal still available in the JCR-psych-d. Twenty-nine such replacements were required.
    - In order to identify non-empirical journals, T.E.H. manually examined each journal's website. If the website indicated that the journal only published non-empirical content (e.g., news, opinions, narrative reviews, systematic reviews,

meta-analyses) it was considered a non-empirical journal and excluded. If the journal's status was not clear from the journal website, T.E.H. examined the last ten articles published in the journal and if they were all non-empirical the journal was excluded.
    - One additional non-empirical journal ("Psychological Science in the Public Interest") in the high-impact sample was identified during data extraction and replaced.
- When journals included in the random sample appeared in the high impact sample, they were removed from the high impact sample and replaced. Four such replacements were required.
- If journals in the high impact sample appeared in multiple subfields, the highest ranked entry was retained and the lowest ranked entry replaced. Three such replacements were required.

Sampling procedure

The random sample was obtained first. We used R to randomly sample 40 journals from the WOS-psych-d, applied the inclusion and exclusion criteria (see above), and randomly sampled additional replacement journals as necessary.

The high-impact sample was obtained second. We used R to select the top-5 journals in each subject area of the JCR-psych-d, applied the inclusion and exclusion criteria (see above), and selected additional replacement journals from the next available rank in the relevant subfield as necessary.

Assignment to coders

Journals in both samples for which there was not already data in TOP-d (see "Obtaining extant data and bibliographic information") were randomly assigned to T.E.H. and B.A.N. such that each coder was assigned to half of the available journals in each sample.

Procedure

1. Each coder completed the data extraction form (https://osf.io/fkvsp/) for each journal assigned to them in the high impact sample (https://osf.io/jyrh8/) and the random sample (https://osf.io/8qzhm/).
2. Information about journal policies was obtained by manual inspection of journal websites. Particular attention was paid to webpages related to instructions to authors and editorial policy. All examined webpages were preserved using the perma.cc service (https://perma.cc/; links are provided in the data files).
3. Completed data extractions were submitted to TOP-d and passed a limited quality control assessment performed by the stewards of that database. This process is not standardized and cannot be considered an independent data extraction exercise.

Supplementary results

Table S9. Counts, percentages, and 95% confidence intervals for the random sample.

policy	policy level (n, %)			
	0	1	2	3
Data citation	29 (72% [60,86])	10 (25% [12,38])	1 (2% [0,16])	0 (0% [0,13])
Data transparency	33 (82% [72,93])	5 (12% [2,23])	2 (5% [0,16])	0 (0% [0,11])
Analysis transparency	35 (88% [80,98])	4 (10% [3,20])	1 (2% [0,13])	0 (0% [0,10])
Materials transparency	35 (88% [80,98])	4 (10% [3,20])	1 (2% [0,13])	0 (0% [0,10])
Reporting guidelines	30 (75% [65,89])	7 (18% [7,32])	3 (8% [0,22])	0 (0% [0,14])
Preregistration (study)	36 (90% [82,98])	2 (5% [0,13])	2 (5% [0,13])	0 (0% [0,8])
Preregistration (analysis)	36 (90% [82,98])	2 (5% [0,13])	2 (5% [0,13])	0 (0% [0,8])
Replication	35 (88% [80,97])	2 (5% [0,15])	0 (0% [0,10])	3 (8% [0,17])
Publication bias	33 (82% [72,94])	0 (0% [0,11])	0 (0% [0,11])	7 (18% [7,29])
Badges	33 (82% [72,93])	1 (2% [0,13])	6 (15% [5,26])	N/A[1]

[1] There is no data in this cell because TOP does not specify a level 3 for the open badges policy.

Table S10. Counts and percentages for the high-impact journals sample.

policy	policy level (n, %)			
	0	1	2	3
Data citation	29 (58%)	17 (34%)	3 (6%)	1 (2%)
Data transparency	35 (70%)	11 (22%)	4 (8%)	0 (0%)
Analysis transparency	41 (82%)	8 (16%)	1 (2%)	0 (0%)
Materials transparency	42 (84%)	6 (12%)	2 (4%)	0 (0%)
Reporting guidelines	28 (56%)	14 (28%)	8 (16%)	0 (0%)
Preregistration (study)	43 (86%)	5 (10%)	2 (4%)	0 (0%)
Preregistration (analysis)	42 (84%)	5 (10%)	2 (4%)	1 (2%)
Replication	36 (72%)	8 (16%)	0 (0%)	6 (12%)
Publication bias	43 (86%)	0 (0%)	0 (0%)	7 (14%)
Badges	43 (86%)	1 (2%)	6 (12%)	N/A[1]

[1] There is no data in this cell because TOP does not specify a level 3 for the open badges policy.

Open practices statement

The study protocol (rationale, methods, and analysis plan) was pre-registered on the Open Science Framework on 4th September, 2020 (https://osf.io/n9325/). There were no deviations from the original protocol. All data, materials, and analysis scripts related to this study are publicly available on the Open Science Framework (https://osf.io/jf7mn/). To facilitate reproducibility, a reproducible version of the analyses are available in a Code Ocean container (https://doi.org/10.24433/CO.4977248.v1) which re-creates the software environment in which the original analyses were performed.

Conflict of interest statement

B.A.N. is co-founder and Executive Director of the Center for Open Science and was involved in the development, promotion, and administration of the Transparency and Openness Promotion Guidelines. T.E.H. declares no conflict of interest.

Changes in job advertisements

We examined whether research institutions are explicitly communicating expectations for replicability and transparency in their job advertisements. We analyzed the entire set of academic job offers in psychology kindly provided from the German platform *academics.de* (provided by the ZEIT newspaper). This is one of the most comprehensive platforms for academic job offers across disciplines in German-speaking countries. We restricted the selection to the existing category “Psychologie, Psychotherapie”. As we were only interested in academic job offers, we furthermore restricted the area to “Lehre & Forschung, Wissenschaft, Forschung & Entwicklung” and job position to “Postdoc, Assistent/in, Referent/in, Wissenschaftliche/r Mitarbeiter/in, Forscher/in, Dozent/in, Lecturer, Lehrkraft, Lehrbeauftragte/r, Studienleiter/in, Akademische/r Rat/Rätin, Studienrat/-rätin”.

Sample Descriptives

The database search returned 1626 job ads (1484 in German, 142 in English). The available ads have been published between 2017-02-07 and 2020-12-29. Table S11 shows the distribution of advertised positions (top 8 categories; other categories have < 1% each and are collapsed in “Other”).

Table S11: Percentage of job ads for different positions

Position	**%**
Professor/in	42
Wissenschaftliche/r Mitarbeiter/in	17
Doktorand/in;Wissenschaftliche/r Mitarbeiter/in	15
Other	13
Postdoc;Wissenschaftliche/r Mitarbeiter/in	4
Dozent/in, Lecturer, Lehrkraft, Lehrbeauftragte/r, Studienleiter/in	3
Gruppen-, Team-, Labor-, Abteilungsleitung	2

Postdoc	2
Professor/in;Psychologe/-in	2

Analysis of open science criteria

Main texts of the job ads were searched for the following (partly truncated) keywords, covering German and English keywords:

```
"open science"          "open-science"          "open source"
"open-source"           "replikation"           "replication"
"replizier"             "reproduzier"           "Reproduktion"
"Reproduzierbarkeit"    "reproducib"            "open data"
"offene daten"          "präregistrier"         "prereg"
"forschungstransparenz"  "transparente forschung" "research transparency"
```

Overall, 2.2% (n=36) of job offers mentioned replicability and transparency as desired or essential job criteria. All of the automatically matched job ads were manually checked that the keyword actually indicated a valid mention of open science. Most of these mentions (n=26) concerned professorship positions, the remainder (n=10) other scientific personnel. Out of 376 advertising institutions, 20 mentioned replicability and transparency at least once. Table S12 shows the fraction of ads mentioning open science split by year.

Table S12: Percentage of ads mentioning open science per year

Year	**%**
2017	0.99
2018	0.99
2019	2.02
2020	3.81

Supplementary Information References

Aczel, B., Kovacs, M., Bognar, M., Palfi, B., Hartanto, A., Onie, S., & Evans, T. R. (2019). *Is there evidence for cross-domain congruency sequence effect? A replication of Kan et al. (2013)*. PsyArXiv. https://doi.org/10.31234/osf.io/5k8rq

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., … Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. https://doi.org/10/bszcmm

Baumeister, R. F., & Vohs, K. D. (2016). Misguided Effort With Elusive Implications. *Perspectives on Psychological Science*, *11*(4), 574–575. https://doi.org/10/gf5srq

Beaudry, J. L., Kaufman, J., Johnstone, T., & Given, L. (2019). *Swinburne Open Science Survey (2019)*. https://lens.org/162-491-164-673-089

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's Threat to Organizational Research: Evidence From Primary and Meta-Analytic Sources. *Personnel Psychology*, *69*(3), 709–750. https://doi.org/10.1111/peps.12111

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-

Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., Goeschl, T., Guo, L., Hauser, O. P., … Wollbrant, C. E. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Brainard, J. (2018). What a massive database of retracted papers reveals about science publishing's 'death penalty.' *Science*. https://doi.org/10.1126/science.aav8384

Brand, A., & Bradley, M. T. (2016). The Precision of Effect Size Estimation From Published Psychological Research: Surveying Confidence Intervals. *Psychological Reports*, *118*(1), 154–170. https://doi.org/10.1177/0033294115625265

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M., Hamlin, J. K., Kline, M., Kominsky, J., Kosie, J., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349–363. https://doi.org/10.1037/cap0000216

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818. https://doi.org/10.1037/0022-3514.45.4.805

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson,

S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. https://doi.org/10/f7mgzn

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*. https://doi.org/10/gckfrd

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., … Yong, J. C. (2016). Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. https://doi.org/10/gf6752

Christensen, G., Wang, Z., Paluck, E. L., Swanson, N., Birke, D. J., Miguel, E., & Littman, R. (2019). *Open Science Practices are on the Rise: The State of Social Science (3S) Survey*. https://doi.org/10.31222/osf.io/5rksu

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Colling, L. J., Sz\Hucs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., & Schroeder, P. A. (2020). Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003). *Advances in Methods and Practices in Psychological Science*, 2515245920903079.

Collins, H. (1992). *Changing order: Replication and induction in scientific practice*. University of

Chicago Press.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2018). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*. https://doi.org/10/gf28qh

Cunningham, M. R., & Baumeister, R. F. (2016). How to Make Nothing Out of Something: Analyses of the Impact of Study Sampling and Statistical Interpretation in Misleading Meta-Analytic Conclusions. *Frontiers in Psychology*, *7*. https://doi.org/10/gfc5jq

CWTS. (2017). *Open data report*. https://www.elsevier.com/about/open-science/research-data/open-data-report

Dang, J. (2016). Commentary: A Multilab Preregistered Replication of the Ego-Depletion Effect. *Frontiers in Psychology*, *7*. https://doi.org/10/f9scg9

Dang, J. (2018). An updated meta-analysis of the ego depletion effect. *Psychological Research*, *82*(4), 645–651. https://doi.org/10/f94cjd

Dang, J., Barker, P., Baumert, A., Bentvelzen, M., Berkman, E., Buchholz, N., Buczny, J., Chen, Z., De Cristofaro, V., de Vries, L., Dewitte, S., Giacomantonio, M., Gong, R., Homan, M., Imhoff, R., Ismail, I., Jia, L., Kubiak, T., Lange, F., … Zinkernagel, A. (2021). A Multilab Replication of the Ego Depletion Effect. *Social Psychological and Personality Science*, *12*(1), 14–24. https://doi.org/10/ggtpft

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., Chung, L. Y., Hermann, A. D., Joy-Gaba, J. A., Line, M. J., Rule, N. O., Sacco, D. F., Vaughn, L. A., & Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology*, *69*, 184–

186. https://doi.org/10.1016/j.jesp.2016.12.005

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Fountain, J., & Harrison, G. W. (2011). What do prediction markets predict? *Applied Economics Letters*, *18*(3), 267–272.

Fox, N., Honeycutt, N., & Jussim, L. (2018). *How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users.* https://doi.org/10.31234/osf.io/3v7hx

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Ghelfi, E., Christopherson, C. D., Urry, H. L., Lenne, R. L., Legate, N., Ann Fischer, M., Wagemans, F. M. A., Wiggins, B., Barrett, T., Bornstein, M., de Haan, B., Guberman, J., Issa, N., Kim, J., Na, E., O'Brien, J., Paulk, A., Peck, T., Sashihara, M., … Sullivan, D. (2020). Reexamining the Effect of Gustatory Disgust on Moral Judgment: A Multilab Direct Replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science*, *3*(1), 3–23. https://doi.org/10.1177/2515245919881152

Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *ArXiv:2102.00517*. https://arxiv.org/abs/2102.00517

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology*, *67*(4), 627.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hardwicke, Tom E., Szucs, D., Thibault, R. T., Crüwell, S., Van den Akker, O., Nuijten, M. B., & Ioannidis, J. P. A. (2021). *Post-replication citation patterns in psychology: Four case*

studies.

Hardwicke, Tom Elis, Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M., & Ioannidis, john. (2020). *Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017)* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/9sz2y

Heathers, J. A., Anaya, J., van der Zee, T., & Brown, N. J. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)* [Preprint]. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26968v1

Héroux, M. E., Loo, C. K., Taylor, J. L., & Gandevia, S. C. (2017). Questionable science and reproducibility in electrical brain stimulation research. *PLOS ONE*, *12*(4), e0175635. https://doi.org/10.1371/journal.pone.0175635

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285.

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Janke, S., Daumiller, M., & Rudert, S. C. (2019). Dark Pathways to Achievement in Science: Researchers' Achievement Goals Predict Engagement in Questionable Research Practices. *Social Psychological and Personality Science*, *10*(6), 783–791. https://doi.org/10.1177/1948550618790227

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10/f33h6z

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R.,

Christopherson, C. D., Clay, S., Collisson, B., & Crawford, J. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. https://doi.org/10/ghwq2w

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kossmeier, M., Vilsmeier, J., Dittrich, R., Fritz, T., Kolmanz, C., Plessen, C. Y., Slowik, A., Tran, U. S., & Voracek, M. (2019). Long-Term Trends (1980–2017) in the *N* -Pact Factor of Journals in Personality Psychology and Individual Differences Research. *Zeitschrift Für Psychologie*, *227*(4), 293–302. https://doi.org/10.1027/2151-2604/a000384

Krishna, A., & Peter, S. M. (2018). Questionable research practices in student final theses – Prevalence, attitudes, and the role of the supervisor's perceived attitudes. *PLOS ONE*, *13*(8), e0203470. https://doi.org/10.1371/journal.pone.0203470

Laudan, L. (1981). Peirce and the Trivialization of the Self-Corrective Thesis. In L. Laudan (Ed.), *Science and Hypothesis: Historical Essays on Scientific Methodology* (pp. 226–251). Springer Netherlands. https://doi.org/10.1007/978-94-015-7288-0_14

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. (2018). Self-Esteem, Self-Disclosure, Self-Expression, and Connection on Facebook: A Collaborative Replication

Meta-Analysis. *Psi Chi Journal of Psychological Research*, *23*(2), 98–109. https://doi.org/10.24839/2325-7342.JN23.2.98

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183.

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. (2019). *Questionable and Open Research Practices in Education Research* [Preprint]. EdArXiv. https://doi.org/10.35542/osf.io/f7srb

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, *91*(3), 425–429. https://doi.org/10.1016/j.econlet.2006.01.004

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10/ghwq2t

Marsh, A. A., Rhoads, S. A., & Ryan, R. M. (2019). A multi-semester classroom demonstration yields evidence in support of the facial feedback effect. *Emotion*, *19*(8), 1500. https://doi.org/10/gf5sj2

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample Size in Psychological Research over the Past 30 Years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487. https://doi.org/10/f7qwxd

Mccarthy, R., Gervais, W., Aczel, B., Al-Kire, R., Baraldo, S., Baruh, L., Basch, C., Baumert, A., Behler, A., Bettencourt, A., Bitar, A., Bouxom, H., Buck, A., Cemalcilar, Z., Chekroun, P., Chen, J., Díaz, Á., Ducham, A., Edlund, J., & Zogmaister, C. (2020). A Multi-Site Collaborative Study of the Hostile Priming Effect. *Collabra Psychology*.

McCarthy, R. J., Hartnett, J. L., Heider, J. D., Scherer, C. R., Wood, S. E., Nichols, A. L.,

Edlund, J. E., & Walker, W. R. (2018). An Investigation of Abstract Construal on Impression Formation: A Multi-Lab Replication of McCarthy and Skowronski (2011). *International Review of Social Psychology*, *31*(1), 15. https://doi.org/10.5334/irsp.133

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., Acar, O. A., Aczel, B., & Bakos, B. E. (2018). Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*(3), 321–336.

McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, J. E. (2021). *Self-Correction in Psychological Science: How do Psychologists Update their Beliefs in Response to Replications?*

Moran, T., Hughes, S., Hussey, I., Vadillo, M. A., Olson, M. A., Aust, F., Bading, K. C., Balas, R., Benedict, T., & Corneille, O. (2020). *Incidental attitude formation via the surveillance task: A registered replication report of Olson and Fazio (2001)*. https://doi.org/10/ghwq2z

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*(1), 34–58. https://doi.org/10.1037/pspa0000084

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*(5), 657. https://doi.org/10/gddggv

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237. https://doi.org/10.1177/2515245920918872

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., Alshaif, N., Andringa, R., Aveyard, M., Babincak, P., Balatekin, N., Baldwin, S. A., Banik, G., Baskin, E., Bell, R., Białobrzeska, O., Birt, A. R., Boot, W. R., Braithwaite, S. R., … Zrubka, M. (2018). Registered Replication Report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268–294. https://doi.org/10.1177/1745691618755704

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole et al.(2016). *Journal of Experimental Social Psychology*, *67*, 86–87.

Plott, C. R., & Sunder, S. (1988). Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets. *Econometrica*, *56*(5), 1085–1118. https://doi.org/10.2307/1911360

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. (2020). *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable*. PsyArXiv. https://doi.org/10.31234/osf.io/n2a9x

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology*, *128*(6), 493–499. https://doi.org/10.1037/abn0000435

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646–656. https://doi.org/10.1037/0022-006X.58.5.646

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, *23*, 114–122. https://doi.org/10.1016/j.psychsport.2015.11.005

Sedlmeier, P., & Gigerenzer, G. (1992). *Do studies of statistical power have an effect on the power of studies?*

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Strack, F. (2016). Reflection on the Smiling Registered Replication Report. *Perspectives on Psychological Science*, *11*(6), 929–930. https://doi.org/10.1177/1745691616674460

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768. https://doi.org/10/dhnv53

Swift, J. K., Christopherson, C. D., Bird, M. O., Zöld, A., & Goode, J. (2020). Questionable research practices among faculty and students in APA-accredited clinical and counseling

psychology doctoral programs. *Training and Education in Professional Psychology*. https://doi.org/10.1037/tep0000322

Tierney, W., Hardy, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., & Uhlmann, E. L. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes*, *161*, 291–309. https://doi.org/10.1016/j.obhdp.2020.07.002

Van den Akker, O., Scherer, L. D., Wicherts, J. M., & Koole, S. (2020). *Support for Open Science Practices in Emotion Science: A Survey Study* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ub4wc

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra*, *1*(1). https://doi.org/10.1525/collabra.13

Vassar, M., Jellison, S., Wendelbo, H., & Wayant, C. (2020). Data sharing practices in randomized trials of addiction interventions. *Addictive Behaviors*, *102*, 106193. https://doi.org/10.1016/j.addbeh.2019.106193

Vazire, S., & Holcombe, A. O. (2020). *Where Are The Self-Correcting Mechanisms In Science?* PsyArXiv. https://doi.org/10.31234/osf.io/kgqzt

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., & Bakos, B. E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, *1*(3), 299–317.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A., Wagenmakers, E.-J., & Albarracin, D. (2021). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L.,

Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, Eric-Jan, & Gronau, Q. (2018, May 10). Musings on Preregistration: The Case of the Facial Feedback Effect. *Bayesian Spectacles*. https://www.bayesianspectacles.org/musings-on-preregistration/

Wagge, J., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., & Grahe, J. (2018). *A Demonstration of the Collaborative Replication and Education Project: Replication Attempts of the Red-Romance Effect*. PsyArXiv. https://doi.org/10.31234/osf.io/chax8

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L. J., Yantis, C., Wong, K. M., Sun, J., Prims, J. P., Mueller, A. B., Melton, Z. J., & Carsel, T. S. (2018). Why Do Some Psychology Researchers Resist Adopting Proposed Reforms to Research Practices? A Description of Researchers' Rationales. *Advances in Methods and Practices in Psychological Science*, *1*(2), 166–173. https://doi.org/10.1177/2515245918757427

Wintle, B. C., Fraser, H., Singleton-Thorn, F., Hanea, A. M., Wilkinson, D. P., Bush, M., McBride, M., Gould, E., Head, A., Rumpff, L., & Fidler, F. (2021). *Eliciting reasoning about replicability in social and behavioural sciences.*

Wolff, W., Baumann, L., & Englert, C. (2018). Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLOS ONE*, *13*(6), e0199554. https://doi.org/10.1371/journal.pone.0199554