**Citation in APA format:** Giner-Sorolla, R. (2025). Changing practices and priorities in social psychological research methods and reporting In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed.). Situational Press.
https://doi.org/10.70400/ZUTF8520

# Changing Practices and Priorities in Social Psychological Research Methods and Reporting

**Roger Giner-Sorolla**, University of Kent

Social psychology, as a sub-discipline of psychology, has been affected by four major methodological transformations in its relatively short lifetime. The first, around 1910, happened when Titchener and Boring steered psychology in the United States away from applied purposes into an experimentalist, theory-focused discipline (O'Donnell, 1979). This change elevated basic research over applied research, an inequality that persisted for many decades (Giner-Sorolla, 2019). The second can be credited to Floyd Allport (Greenwood, 2000), who established an individual-based social psychology in the 1920's, in contrast to collective levels of explanation which continued to find a home in the social psychology that is a sub-discipline of sociology. The third transformation, in the late 1960's and 1970's, questioned social psychology's relevance to increasingly visible social problems, as well as the ethics and validity of the standard model of laboratory research (Elms, 1975). Other social psychologists, especially in Europe, pushed back against the individualist and experimentalist order established in a previous era, reviving and further developing collective and discursive approaches (Hogg & Williams, 2000). Fourth, those decades also saw worries about whether a single-study article with confidently stated conclusions based on a significance criterion of .05 made an adequate basis for evidence (e.g., Lykken, 1968; Meehl, 1978).

Changes to evidence standards in the 1980's and 1990's did not completely quell these last worries. But in the past decade, concerns about the robustness of evidence and transparency of reporting , some long-standing and others relatively new, have started to be heard by editors, professional societies, and other gatekeepers in social psychology. As a result, social psychology has seen widespread changes to methods and reporting standards in recent years. It is this time of shifting values that must be captured, more than any technical advance in research methods, by a chapter being written to update a handbook whose last edition was published in 2010. But to evaluate why the latest innovations have been adopted, we must understand the way things used to work.

# THE SYSTEM OF RESEARCH AND REPORTING, CIRCA 2010

The system for publishing original research in social psychology involves deciding both how valuable the research ideas are, and how strong the empirical evidence is for the claims being made. In both judgments there is a tension between easy-to-apply, consistent rules that may not capture their construct accurately, and complex judgments that promise more nuance alongside more controversy, because different people may come to different decisions. In a time of fresh consideration of statistics and methodology in our field, when rules are changing or becoming more nuanced, it is easy to take a cynical view of people who followed the old rules. However, rather than paint most pre-2010 research as lazy at best, or dishonest at worst, I take the view that most researchers trying to get published genuinely aligned their idea of good research to that system, while the standards and rules of the new system are not beyond question. Both assumptions shape my further review of the issues.

What, then, were the elements of research methodology and reporting that prevailed in mainstream social psychology since the 1980's, with few challenges until recently? These elements were not completely described by the norms for writing psychology articles set forth, for example, by the American Psychological Association in its style manual and other documents (e.g. American Psychological Association, 2010). Some of their assumptions, however, can be gleaned from less formal guides to research, writing, article reading, and statistical inference in psychology (e.g. Abelson, 1995; Bem, 2003; Cook & Groom, 2003; Jordan & Zanna, 1999; Ritter et al., 2012; Sternberg & Sternberg, 2010).

The system both now and then has asked researchers to present a series of empirical studies that support a hypothesis, then sets out norms for evaluating these studies. Success requires anticipating and meeting these evaluative norms in the competitive world of prestigious journals, where no more than 20% of submitted articles are accepted (see Bakker, 2012, Giner-Sorolla, 2012, and Schneider, 1992). Under such pressure, research in social psychology has been historically rewarded for four features that weigh in the judgments of editors and reviewers.

- **Strength of evidence.** After a period of doubt in the 1970's about how much could be concluded from a single study with a significant finding, multi-study papers increasingly became the norm in social-personality psychology (Quiňones-Vidal et al., 2004; Reis & Stiller, 1992; R. C. Sherman et al., 1999). Even with multiple studies, selective pressure continued to work. An informal norm soon evolved: to avoid the article's evidence being rejected as weak or inconsistent, every study's key null-hypothesis test must show a finding, significant at $p < .05$. Contemporary documentation of this process is hard to obtain, but it corresponds to personal and colleagues' experience (documented to some extent by Greenwald, 1975). It also fits with studies of the psychology literature in which, consistently, more than 95% of reported results support the hypothesis (Bakker et al., 2012; Motyl et al., 2017; Scheel et al., 2021). This is a higher rate of confirmation than would be expected even if all hypotheses were true and tested by reasonably well-powered studies (Francis, 2017). One more exception rarely acknowledged in bibliometric studies may have involved p-values in a "marginal" range from .05 to .10 (Pritschet et al., 2016). Marginal p-values were sometimes allowed to stand as positive evidence in a single study, provided other studies in the package were conventionally significant. Thus, even if a few nonsignificant results appeared in multi-study papers, some of these might have been tolerated only because they were approaching significance.

- **Novelty**. Repeatedly and overtly, guides to academic careers, writing, and journal submission pre-2010 emphasized the value of establishing something new (for a summary of novelty as a standard in psychology journals, see Berkman & Wilson, 2021). At the highest levels of prestige, novelty is expected to be generative: a new framework, hypothesis, or phenomenon. But at a minimum, research is expected to show combinatorial novelty: for example, the interaction of a manipulation and an individual difference which have each been investigated separately. Less well regarded were conceptual replications, testing an already-shown relationship between constructs using different manipulations or measures. Direct replications, despite intermittent efforts to create channels for their dissemination (see Giner-Sorolla, 2012), also did not attract much interest. Cumulative and incremental research, in this hierarchy of value, takes second place to research that claims a more novel advance.
- **Simplicity.** Guides to writing often counsel keeping the story simple: "A journal article […] is not a novel with subplots, flashbacks, and literary allusions, but a short story with a single linear narrative line" (Bem, 2003, p. 173). This advice extends to data reporting: "Think of your dataset as a jewel. Your task is to cut and polish it, to select the facets to highlight, and to craft the best setting for it." (p. 173). In the backlash against this advice, Bem's accompanying caveat is sometimes forgotten: "The integrity of the scientific enterprise requires the reporting of disconfirming results" (p. 175). Likewise, APA publication manuals for decades have proclaimed "Psychologists do not omit troublesome observations from their reports so as to present a more convincing story." (e.g. American Psychological Association, 1994, p. 12, and repeated in every edition since then). Despite these statements of ethical principle, increased competition over the years meant that authors felt unable to present a narrative showing any inconclusive results. The editor, at best, might ask the authors to simplify the narrative, and at worst would reject the paper. Articles in the most prestigious journals describing theoretically grounded research with no consistent pattern of results were rare enough in an earlier era[1], and almost inconceivable twenty years later.
- **Sophistication.** Simplicity, however, had a limit, because merely descriptive reports didn't fare well either. In this system of evaluation, making a story out of descriptive data required a certain level of explanatory sophistication. As Jordan and Zanna (1999) advised writers, "Facts in isolation are trivia. Facts tied together by an explanatory theory are science. Therein lies the story" (p. 467). Simple demonstrations were expected to be followed up in the same article by more elaborate designs establishing a process. Likewise, experiments supporting a cause-and-effect narrative have been normatively favored over descriptive and correlational results, particularly in social psychology (Bless & Burger, 2016; Li et al., 2012; R. C. Sherman et al., 1999; Tracy et al., 2009). By the 21st century, research increasingly has had to show third-variable processes (e.g., through mediation analysis) or context effects (through moderation), a trend that has attracted pushback. Cialdini (2009), for example, complained that increasing requirements for sophisticated analyses disadvantage the simpler experimental designs necessary to field research. Rozin (2001) likewise argued for increasing acceptance of descriptive research.
- **Generality.** Articles were expected to emphasize their implications for wider knowledge. Note, in the above quote from Jordan and Zanna (1999), that what is sought is not just a story, but a theory. Many journals also did, and still do, require theoretical relevance in their submission guidelines. Relying on theory goes together with the de-emphasizing of particular traits of the sample, setting, or method. Findings are meant to be presented not just as a case study, but as the first evidence in a general model of human psychology.

In highly competitive environments, a single weakness is fatal (Bornmann & Marx, 2012). High pressure, then, has provided strong incentives for "good science" to be defined in terms of a product that shows each of the desired characteristics. Mixed packages that are strong on some desirables, but deficient in others, are not likely to survive in a high-pressure system. Yet if the norms are

followed consistently, so that authors know what to expect, the system's existence benefits researchers and evaluators alike. It produces a consensus within a field, a process whose rules can be applied in a transparent and consistent way. As Gregg and Sedikides write (2004), "The purpose of [null-hypothesis significance testing] is to set standard criteria –collectively agreed upon by members of the scientific community – that must be met by any putative effect before it can be provisionally admitted into the Pantheon of the Real." (p. 340).

To support these criteria of research value, social psychology has adopted a common version of statistics, handed down from advisor to student, and reinforced through the interventions of peer review (Gigerenzer et al., 1989; Goodman, 2019). While this good-enough version may sometimes be overly simplistic or wrong in the view of statisticians (see Fife, 2020), it smooths the task of doing research by ensuring a common language within a single discipline. But at times, the common version needs updates to remedy visible failings. At the same time, it should avoid faddish innovations such as "$p_{rep}$", a proposed replacement for p-values that added little value over usual practice (Iverson et al., 2009). In updating researchers according to the best advice of statisticians, the role of the "maven" who understands the standards and needs of both groups may very well be essential (Sharpe, 2013).

Simulations varying incentives and output in a computer model of the culture of science have illustrated how commonly held methodological standards can cling on tenaciously, sometimes with negative effects (Bakker et al., 2012; Smaldino & McElreath, 2016; Wu et al, in press). However, standards can evolve, as when the field of social psychology wholesale adopted mediation analysis in the 1990's. Often, change happens in response to specific pressures within the system. In the case of mediation, sophistication was encouraged by editors who judged that a three-variable process narrative was more informative than a mere two-variable effect. Other times, standards evolve due to external pressures, most urgently when a self-evident failure case impugns the old ways. For example, in the 1970's, classic studies such as the Milgram conformity experiments and Zimbardo's prison experiment came into question for their flawed ethics of consent and deception. The questioning resulted in a long-lasting change in research culture, improving ethical procedures while shifting away from highly involving situations, for better or worse (West & Gunn, 1978; Vitelli, 1988).

Consensus and consistency are also helped by adopting heuristics -- simple decision rules such as "p < .05 means a conclusion is valid." The five criteria I have mentioned can be expressed as heuristics: *evidence,* "Is the claim supported by multiple significant results?"; *novelty*, "Are the ideas and findings shown to be new?"; *simplicity*, "Does the research have a clear message that supports a priori hypotheses?"; *sophistication*, "Does the research tell a story rather than just set a scene?"; and *generality*, "Is the contribution to knowledge understood to go beyond the particulars of the setting studied?"

All heuristic rules, though, carry the risk that, eventually, superficial signs of compliance will be used to meet them without meeting their underlying standards. This risk holds true in academic publishing, just as it does in social programs, the original context of the observation known as "Campbell's law" (Campbell, 1979). And, as observed in the corollary "Goodhart's law" from economics, the urge to take shortcuts becomes more tempting when a system is put under pressure (Goodhart, 1975; see Fire & Guestrin, 2019, for an application to academic publishing). Specifically, each of the heuristics listed above has a less effortful shortcut that has undermined its accuracy in the past. Unless the easier ways are blocked by the conscientiousness of informed authors, reviewers, and editors, they will win because of their greater efficiency.

- Shortcuts to **evidence**: choose one out of many analytic methods to make the result look stronger, as in the practice of "p-hacking" (Simonsohn et al., 2011); or, selectively report only supportive studies among the ones actually conducted (e.g., Ioannidis et al., 2014; van der Steen et al., 2018); or, more trivially, use one-tailed significance tests without meeting their theoretical preconditions (Goldfried, 1959) to give the superficial appearance of "$p < .05$" to results that are $p < .10$, two-tailed.
- Shortcuts to **novelty**: repackage existing ideas using new terms and psychometrically near-identical measures (Hodson, 2021; Kelley, 1927).
- Shortcuts to **simplicity**: misrepresent the actual complexity of the data by selectively reporting story-congruent results; or, streamline the story by passing off exploratory results as predicted a priori (Kerr, 1998); or, avoid appropriate but unfamiliar analytic techniques because, ironically, reviewers might reject them as just another way of gaming the data.
- Shortcut to **sophistication**: make rote use of advanced techniques without considering their fit to the assumptions of the data, or their appropriateness for the question – as when mediation analyses are included because of their resemblance to a process model, which has often been accepted by the norms of the field even though the variables are causally ambiguous or not conceptually distinct from each other (Bullock et al., 2010; Fiedler et al., 2011, 2018; Spencer et al., 1995). Presenting an underspecified claim where a fully developed theory is needed is another common shortcut to sophistication (Scheel, 2022).
- Shortcut to the appearance of **generality**: use participant types and settings (such as US American undergraduate students in the laboratory) that do not call attention to their particularity (Henrich et al., 2010); or, present generally stated hypotheses without taking on the difficult parts of theory, such as the need to precisely define constructs, auxiliary hypotheses, and boundary conditions (e.g., Fiedler, 2017; Scheel, 2022), and to acknowledge relevant theory outside one's own laboratory (McPhetres et al., 2021).

While heuristics can favor appearance more than substance, there are also benefits to having them. They make decisions easier, and they reinforce consensus about decisions. In systems where context is allowed to modify heuristics, accuracy and sensitivity increase, even as consensus and consistency decrease. Nonetheless, some authors have argued for flexibility in even the most basic decision rules (e.g., Lakens et al., 2018 with p-value criteria). However, others have argued for keeping heuristics, increasing confidence in findings by changing heuristics' parameters (e.g., moving to a $p < .005$ criterion; Benjamin et al., 2018). Both approaches have trade-offs, but if heuristic rules are used, they should be explicitly stated and consciously justified. They should be applied with a good understanding of the conditions under which they are valid. And they should be subject to periodic review as new understanding emerges.

The costs of potential reforms also bear considering. If we abandon existing heuristics or introduce new constraints on research, the process of doing and evaluating research can become more effortful. While the goal of accuracy is laudable, it is necessary to ask how any additional labor cost is going to be met. The system we currently inhabit is one where researchers need to maximize their output to stay productive, and where evaluators are typically uncompensated for their time and increasingly hard to find (e.g. Petrescu & Krishen, 2022). Without tangible career incentives, just exhorting people to improve their practices will have a hard time getting widespread results. It is also worth noting that many recent incentives have been made possible by the development of online technology to reduce their costs and limitations. For example, complete reporting has been eased by the availability of online supplementary material to circumvent the word limits of paper journals. Accessible preregistrations, too, would also not be possible without the Internet.

With cost-benefit tradeoffs in mind, some scholars have used economic simulations, using knowledge rather than material goods as output, to address how scientists should maximize the accuracy and utility of their research given finite resources (Finkel et al., 2017; LeBel et al., 2017a, 2017b; Miller & Ulrich, 2021). In such analyses, incentives for individual scientists may not align with incentives for truth-finding or productivity on a field-wide basis. Misaligned incentives are persuasive explanations for many of the inadequacies that can be shown in a research literature. They will have to be realigned to achieve lasting and widespread remedies for social psychology's methods and reporting problems (Nosek et al., 2012).

In the analysis that follows, I will use the framework of heuristics, shortcuts, costs, and benefits to explain three broadly stated challenges to the way research is done and reported in social psychology. The challenge that has sparked the most change concerns the rules of **evidence** by which sets of findings are judged suitable for publication. I review eight solutions for the problems of the standard, 2010-era system. For each solution, after explaining why it has been advanced, I consider evidence for uptake of the solutions in the past ten years; describe their costs and limitations; and end with ways in which the solutions could lead to new rules for evaluating research. I then apply a parallel analysis to two emerging critical themes that have not seen as much concrete success in changing how journals and researchers do business: challenges to the way we use **methodology**, specifically measurement and manipulations; and challenges to the assumption of **generality** in our research.

# CHALLENGES TO THE RULES OF EVIDENCE: EIGHT PROPOSED SOLUTIONS

In 2011, the heuristics underlying the publication model in social psychology took a direct hit from a controversy that was seen as a failure case for business-as-usual in the field[2]. The Journal of Social and Personality Psychology published a nine-experiment article by the eminent social psychologist Daryl Bem, making the case that effects could be obtained on responses *before* the stimulus was presented: that is, precognition (Bem, 2011). When this result was released, many observers wondered whether the rules of publishing were working well enough to prevent mistaken claims from being presented as fact. Social psychology's highest-impact empirical journal had published a paper that, structurally, looked like any other. It presented an extremely novel claim, with eight studies in support with a significant result, and one marginally significant at $p = .098$. These studies presented variations in method to show that precognition could replicate and generalize, including state-of-the-art reaction-time priming methods. But the conclusion, if valid, would shatter generally accepted understandings of psychology, neurobiology, and indeed physics. In the wreckage of this collision, the standards of evidence that would allow this paper to be published came into question themselves.

The most thorough and immediate critiques of Bem (2011)'s evidence came from LeBel and Peters (2011) and Wagenmakers et al. (2011). With some overlap, these replies pointed out that the paper consisted of many conceptual replications which lacked a systematic process of direct replication or validation. Among a series of conceptual replications, "it becomes easy to attribute negative results to methodological flaws and hence relegate them to the file drawer" (LeBel & Peters, 2011, p. 374). Wagenmakers et al. (2011) likewise observed that the studies garnered their significant results from a collection of analytic methods which were often not replicated within the paper, appearing to confuse exploratory and confirmatory approaches. In one study, precognition appeared only in

response to erotic stimuli; in others, it appeared only among women, or among people scoring high on a sensitivity questionnaire. Indeed, by the later admission of Bem and one of his research assistants, an unrecorded number of failed studies was conducted during the project, which spanned a decade (Engber, 2017). It seemed that, using Bem's own metaphor, too much "cutting and polishing" of the data jewel was going on, and not enough of the ethically mandatory "reporting of disconfirming results."

The second criticism shared by the two articles was a two-pronged appeal for gatekeepers in psychology to consider prior evidence both informally by acknowledging that "extraordinary claims require extraordinary evidence" (Wagenmakers et al., p. 429), and formally by showing that Bayesian statistics would have found little evidence in most of the studies. As it turned out, the formal Bayesian statistics gave a more equivocal result because their criterion was approximately equivalent to $\alpha = .01$. Indeed, that significance criterion would have given a similar negative evaluation of the studies if applied using null-hypothesis statistics (Jeon & de Boeck, 2017). But the larger, informally stated point would stick. Bem's article describes a process in which the average person gets intuitions about random stimuli seconds in advance. Surely millions of casinos couldn't be missing a shortfall in their mathematically predictable house takings, as millions of everyday psychics place last-minute bets on the roulette wheel? The finding was surprising precisely because it did not line up with common experience, scientific understandings of how the mind works, or even with folk beliefs that only rare individuals are gifted with the second sight.

Considerations of prior evidence and plausibility bring controversy. Specifically, they collide with one of the main ideals of the publishing game: to create a level playing field where all ideas are subject to the same null-hypothesis rules. After all, what does it mean to reject a paper because its evidence, although suggestive, is implausible in light of our general worldview? Imagine an editor applying a very low prior to evidence that women are just as intelligent as men are, based on judgments of "plausibility" that are more rooted in prejudice than fact. This sounds dangerously close to enshrining paradigm bias, and maybe even political bias, in our rules of evidence. And yet, the ESP article was so disruptive to the formal assumptions of psychophysics that demanding more evidence would have clearly been appropriate. The difference between priors and prejudice may ultimately come down to a close examination of their roots in evidence.

At about the same time, a more general skepticism of results in the published literature led Leif Nelson, Joseph Simmons, and Uri Simonsohn to publish an article critical of "false-positive psychology" in 2011, which has now been cited over 3,000 times. While the authors have stated for the record that their paper was not a rejoinder to Bem's article (Engber, 1987), their analysis and critique fed into an emerging skeptical zeitgeist. They used simulated sampling from a null-effect population, as well as an actual experiment testing a manifestly bogus premise, in which adult participants' date of birth was, impossibly, the *dependent* variable. These demonstrations meant to show how selective use of analytic techniques could yield significant results good enough for publication but with little factual basis. Importantly, all the "p-hacking" moves identified in the paper were recognizable as more-or-less accepted practices within social psychology: for example, running additional clumps of participants to improve on a near-significant result, or adding covariates to remove random and confounded variance. The problem lay in letting the significant outcome decide post-hoc the process that would eventually be reported, and then failing to correct p-values for the multiple chances being taken in analysis.

It is true that some social psychologists already knew to watch out for evidence that the p-hacking game was being played beyond the limits of plausibility. Even before 2011, a paper in which the

authors measured two covariates in two different studies, but used a different one in each analysis, would have struck most reviewers and editors as a flagrant instance of p-value gaming and would have been questioned. Indeed, Simmons et al.'s headline false-positive result of 65% for null population effects, based on using all available p-hacks indiscriminately, was probably an overstatement. The sets of analyses such a process would produce would have looked unacceptably inconsistent to any reasonable editor or reviewer, especially when put together into a multi-study article. But at the time, the analysis shook the faith of the field in the statistically wrong but commonly believed proposition that our literature's false-positive rate was close to the 5% of the conventional alpha criterion (Hubbard, 2011). Because recent simulation findings show that p-hacking's negative effects intensify at moderate levels of publication bias (Friese & Frankenbach, 2020), both p-hacking and selective reporting were and still are best addressed as one problem.

A wave of replication projects in psychology began soon after 2011 (see Wiggins & Christophersen, 2019) and further shook faith in the usual way of verifying social psychological claims. The rejection by *JPSP* of an initial set of direct replications that failed to confirm Bem's results (Ritchie et al., 2012) may have well been a fair application of the reigning novelty heuristic, under which direct replications could contribute no new positive knowledge. It may also have been a result of the studies' low statistical power (Aldhous, 2011). However, it also looked uncomfortably like the failure to apply a fundamental tenet of science, in which errors are expected to be corrected by replication (Schmidt, 2009). JPSP, soon after, published a more highly powered series of studies failing to replicate Bem (Galak et al., 2012). But at the same time, other researchers began to be curious about whether they could replicate more mainstream areas of the literature. The result was a series of high-profile, high-investment replication projects, some of which sought to generalize across the psychology literature (Open Science Collaboration, 2015; Ebersole et al., 2016; Klein et al., 2018), and others which focused on a single research question (for summaries and analyses, see Baumeister et al., 2023; McShane et al., 2019; and Olsson-Collentine et al., 2020).

These projects opened a controversy over the rate of replicability in social psychology, focusing on the projects' methods of sampling the literature as well as their methodological and analytic criteria for declaring a replication failed or successful (Etz & Vanderkerckhove, 2016; Gilbert et al., 2016; Patil et al., 2016; Simonsohn, 2015). Regardless of the baseline rates, the overall picture was not ideal, especially when comparing social psychology to cognitive psychology. Specifically, the project's headline result that only 25% of social psychology studies had a significant replication result came as a shock to the field. But many of the non-replications were better categorized as inconclusive because they based their power analyses on effect sizes reported in original research without correcting for publication bias (Simonsohn, 2015). Likewise, under a Bayesian analysis, almost none of the replications provided conclusive evidence for a population effect close to zero (Etz & Vanderkerckhove, 2016). Finally, replication studies sometimes failed to replicate or validate manipulation checks, leaving findings about the outcome inconclusive (Baumeister et al., 2023)

However, even when large-scale replication projects supported the conclusion of their originals by showing an overall effect different from zero, the effect size was often attenuated relative to original studies or meta-analyses (Kvarven et al., 2020). This implied that a bias toward publishing positive results was distorting the literature. Further investigations following PhD dissertations (Cairo et al., 2020), and funded proposals (Franco et al., 2016) to their eventual form as published articles, also have found evidence of selective reporting favoring positive results in the psychology publication process.

Thus, it seems that the most conclusive message from these projects is not so much "all of social psychology can't replicate" as "social psychology effects are smaller than they appear in the literature" (see also Fabrigar & Wegener, 2016). Small but significant effect sizes established in replications, such as the d = .12 obtained by Klein et al. (2018)'s group in replicating Bauer et al's (2012) study of consumerism priming (original d = .87), mean one of two things. Either further efforts to study the phenomenon productively would need to invest beyond the usual limits of the typical laboratory budget (e.g., to study d = .12 at 90% power and conventional significance would require almost 1500 participants per group); or, alternatively, more care would need to be taken ahead of time in determining the optimal method for studying effects in varying contexts (Lewis et al., 2022; van Bavel et al., 2016a). This latter caution is particularly relevant to social psychology because of the high cultural and social context dependence of our hypotheses and methods (Gollwitzer & Schwabe, 2022; Inbar, 2016; Stroebe & Strack, 2014; van Bavel et al., 2016b).

The prominence of the ESP and replication failure cases, as well as fallout from the Stapel and other fraud cases, led to talk of a "crisis" in social psychology (e.g., Earp & Trafimow, 2015). Others have extended the crisis to cover psychology at large (e.g., Malich & Munafò, 2022), or science at large (e.g., Jamieson, 2018). Such a dramatic narrative is usual in the history of social psychology. After all, amid the "crisis of 1977," an article was written reminding readers of the "crisis of 1927" (M. A. Lewin, 1977). But this time, crisis talk underscored a widespread perception that something needed to be done to change the rules of the evidence game. Indeed, there is recent evidence that low replication rates, when brought to the attention of the public, can damage trust in psychological science, damage that is difficult to repair simply by offering reasonable explanations (Hendriks et al., 2020; Wingen et al., 2020).

At this writing, many changes have been put in place by journals and other evaluators of research: more changes, in fact, than an observer aware of the fate of previous attempts in the history of psychology might have expected ten years ago. Below, I describe and evaluate eight of these reforms: reporting *effect sizes* and other statistics that help interpret p-values beyond a simple significance decision; allowing reports of research with *individual null results*; allowing reports of research with *summary null results*; conducting and reporting *direct replications*; increasing *statistical power*; reporting research under *full disclosure conditions*; carrying out *preregistration and registered reports* (as well as being more tolerant of *honestly descriptive research*); and *strengthening the role of theory* in research.

## Evidence Solution 1: Report Effect Sizes

### *Why?*

Before the 2010's, there flowed an undercurrent of dissent to the prevailing heuristics of evidence in psychology. The requirement of statistical significance, $p < .05$, came in for special scrutiny. Among many other barbs aimed at the p-value, it was described as a "null ritual" which would not be missed if abandoned completely (Gigerenzer et al., 1989, 2004), a method that "can be inefficient or pathological" compared to Bayesian inference (Lee & Wagenmakers, 2005, p. 666), and a decision process characterized by "cognitive distortions" (Kline, 2004, Ch. 5).

Some of the drawbacks of a bare comparison of *p* with an alpha criterion can be remedied by reporting exact p-values, effect sizes, the statistical power of the test, and/or confidence intervals.

Such reforms to evidence reporting were proposed decades ago. They have since received the approval of such bodies as the Society for Personality and Social Psychology and the American Psychological Association (Greenwald et al., 1996; Kashy et al., 2009; Wilkinson & Task Force, 1999; American Psychological Association, 2003).

In the analysis that follows, I focus on the issue of effect size reporting for brevity's sake. However, similar reports could be made about the increasingly acknowledged importance of exact p-value reporting and confidence intervals, both of which share the goal of effect size reporting: to go beyond whether an effect is significant, and answer the question "But by how much?"

Effect sizes, in the first place, are useful to people trying to aggregate research through meta-analysis, in which effect size is the common currency of a research question that may have been pursued in different ways. Explicit reporting spares meta-analysts the effort and uncertainty of trying to reconstitute effect sizes from reported test statistics.

Second, effect sizes provide an additional basis for evaluating the finding, alongside the p-value. Specifically, effect sizes remedy several misinterpretations that can happen when only p-values are on the menu (Ferguson, 2016; Wilkinson and Statistical Task Force, 1999). Relying on significance can lead to over-interpreting significant effects from a large sample, even if they are not large in terms of raw units or standardized metrics. Conversely, a non-trivial effect size might better inform the tendency to overinterpret non-significant effects as evidence that nothing much is going on. Effect sizes, more generally, help scientific readers gauge the practical possibilities of a discovered effect, and lay readers as well, if presented in easily interpretable units.

## *Uptake*

Societies and journals have largely supported effect size reporting requirements. For example, in 2010, the fifth edition of the APA publication manual upgraded its standards, making clear that "estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals" (American Psychological Association, 2010, p. 33). As the guidelines became more insistent, evidence was accumulating that they had not always been followed in practice (for evidence of low to mediocre rates of effect size reporting in APA and other journals in the 2005-2014 period, see Fritz et al., 2012, 2013; Kashy et al., 2009; Sun et al., 2010; Szucs & Ioannidis, 2017).

Effect size reporting has nonetheless increased since the time of the APA Task Force report. Motyl et al. (2017) found that among critical significance tests in several highly ranked social and personality psychology journals, the rate of effect size reporting increased from 19.2% in 2003-4 to 49.7% in 2013-4, similar to an increase seen in reporting exact p-values. Also, their sample of over 1,000 social-personality psychologists reported on average "often" to "always" reporting effect sizes. This uptake was much higher than the other scientific reforms asked about, such as open data and power analysis (see also Washburn et al., 2018). Moreover, almost 50% said they had increased their reporting of effect sizes in response to discussions about the status of psychological science. More recent archival and survey data will have to be conducted, however, to see if the trend continues upwards to near-universal effect size reporting.

## *Costs And Limitations*

Thirty years ago, effect sizes often had to be computed by hand. But nowadays, they are easy to request with statistical software. Reporting and interpreting them adds only a few words to each analysis. On the editorial side, effect sizes are a minimal addition to the usual statistical numbers that are checked. Problems to be worked out in more detail include which type of effect sizes to report -- for example, standardized or unstandardized effect sizes (Baguley, 2009), or bias-corrected effect sizes in ANOVA (Mordkoff, 2019). Adding effect sizes, nonetheless, has been one of the easiest reforms to put into place.

### *New Rules?*

Checking for effect size reporting should be straightforward for a journal editor, and manuscripts can be sent back with requests to add missing statistics. But reporting should not be confused with having standards for effect sizes themselves. Should manuscripts be rejected for basing conclusions on effect sizes that are below some threshold of practical significance, even if they are shown to be significantly on one side or another of zero? The answer is a resounding "no"! Conclusions from low effect sizes should be qualified, but it is just as useful to know with a high degree of certainty that two variables are weakly related, or that two groups are minimally different, as that strong relationships exist. And practical significance is always relative to the arena of practice. A behavioral increase of 0.1% may look small and hard to capture, but if verified through large-scale research and applied to transactions totaling $1 billion a year, it is worth a million bucks.

## Evidence Solution 2: Report Null Studies Within An Article

### *Why?*

Consider a package of three studies with significant findings in support of a proposition. Now consider a second package with the same three studies, plus one more study with nonsignificant findings pointing in the same direction as the others. Paradoxically, it became expected in our field that editors would judge the second package as *worse* evidence than the first because of its mixed significance, although it likely has statistically stronger joint evidence. The causes, irrationality, and consequences of the perfectionism paradox have been spelled out elsewhere (Giner-Sorolla, 2012). It is consistent with recent findings that evaluators of research underestimate how much a series of studies containing some nonsignificant results supports a hypothesis (van den Akker et al., 2023a). One direct editorial remedy (e.g., Giner-Sorolla, 2016) is to promote a policy in which evidence in support of a position is judged by preponderance, rather than perfection. This policy suggests a formal meta-analytic approach to summarizing the evidence from multiple studies which may vary in their individual conclusions, or mini-meta-analysis (Goh et al., 2016). This approach includes the possibility that a set of individually non-significant studies may add up to an overall significant conclusion.

Being open to individual studies with nonsignificant hypothesis-relevant results removes incentives to p-hack, for example, by using one-tailed testing selectively (Banks et al., 2016). It puts a more representative chunk of the research literature into the hands of meta-analysts. It also has the beneficial side effect of removing doubt about the honesty of reporting overall. Indeed, a set of studies that satisfies the perfectionism heuristic might ironically look too good to be a complete report of research (Schimmack, 2012). That is, the chance that (say) nine studies, no more, no less,

would each get a significant result is only about 13% assuming they have 80% power each to detect a real, nonzero population effect size. If the power is closer to 50% for detecting a defined small-to-medium effect size, as Fraley and Vazire (2014) found to be usual in social psychology journals from 2006 to 2010, nine significant results in a row becomes much less plausible, at 0.2%. Using a variety of methods such as p-curve, tests of insufficient variance, and R-index, analyses of psychology literatures pre- and post-2011 consistently show that too many significant relative to nonsignificant findings are being published, compared to what reporting would look like without selection for positive results (e.g. Francis et al., 2014; Motyl et al., 2017; Nelson et al., 2019; Stanley et al., 2021).

## *Uptake*

There are over 700 citations of Goh et al. (2016) as of mid-2022. The vast majority of these apply the mini-meta-analysis method to substantive research, showing that the aggregation approach has become popular in psychology. However, it is not evident that a more relaxed approach to study-by-study perfectionism is being taken by editors and authors. Motyl et al. (2017) found no real difference between the rate of significant critical tests in highly-ranked social-personality psychology journals in 2003-4 (89%) and 2013-4 (92%). True, these figures mean that some nonsignificant critical results were being reported. But they can't reflect the whole story of the research being done. That would require us to believe that the *average* statistical power to detect population effects in 2003 was, implausibly, around 90%.

Among the researchers in Motyl et al.'s (2017) survey, many also admitted to selectively reporting studies in articles, more so than other questionable research practices. Detailed responses showed that it was considered more acceptable to omit a study that failed in its manipulation checks or other assumptions, than to do so because of editorial pressure, even though such pressure was also often cited as a reason. Editors-in-chief who want to see a more realistic picture of data in their field should thus explicitly encourage their associates to take a more open view of partially supported findings, and let authors know this through editorial communications and journal guidelines (e.g., Giner-Sorolla, 2016). There is evidence that communicating relaxed significance expectations starting in 2016 (along with other innovations such as open science badges) had some positive effects on the robustness and credibility of reporting at the Journal of Experimental Social Psychology (Schimmack, 2020), although rates of positive results were still higher than expected under non-selective reporting.

## *Costs And Limitations*

Reporting more studies and aggregating their results will lead to manuscripts being longer and researchers putting in more work. This cost might be mitigated by recognition of researchers' labor which has already been sunk into conducting and analyzing studies that would not otherwise be published. The downsides for journals in increased page length are mitigated by the increasing availability of online supplementary materials, and by some journals' ongoing policy to waive word limits on reports of methods and results (e.g. Eich, 2014).

Reporting "failed" studies has sometimes been opposed because researchers know the reasons why they failed and they are trivial -- experimenter error, for example. However, it would be a step forward for norms to change so that researchers keep an audit trail for trivial as well as for more consequential failures, even if only in a footnote. No matter how embarrassing it might be to put

one's failures on record, it is important to get outside input on whether a failure was "trivial" or more substantive.

There is also a frequent misunderstanding about the audience for research articles; to paraphrase, the belief that "nobody wants to hear about your odyssey of failures" (e.g., Bem, 2003, and Sternberg & Sternberg, 2010). This might be true of the casual academic reader who doesn't personally work in the author's specific topic or paradigm. Such readers might be better served by a review article (or, let's face it, by just skimming the Introduction and Discussion sections.) However, a researcher working on the same problems as the article's should be intensely interested in what can cause their studies to yield unexpected results. They are also invested in knowing the whole picture of findings, whether successful, failed, or inconclusive, so they can make decisions on what to work on. Here, supplementary material comes to the rescue of the information consumer. It makes sure that peer reviewers and interested readers can fact-check the simpler narrative that the main article presents.

Internal meta-analyses carry some limitations. If performed on sets of studies that are selectively reported, they can exacerbate false impressions from the research (Ueno et al., 2016; Vosgerau et al., 2019). Although internal analyses are a good way to encourage the reporting of all evidence, it does bear repeating that this technique in fact *depends upon* the reporting of all relevant evidence. Given the fluid way in which some labs construct multi-study papers from multiple lines of research, it can be tricky to come up with bullet-proof wording for a laboratory-wise disclosure statement about multiple studies. Nonetheless, a few authors have started making such disclosures on their own (e.g., Woolley & Fishbach, 2017: "We disclose all studies testing our hypotheses and all measures within each study", p. 153.) Making these disclosures a requirement would require a shift in field-wide norms. We would have to expect the package of studies that go into a paper to be determined before, not after, doing the research. Many would find it difficult, given the flexible way in which studies often build on and learn from each other.

### *New Rules?*

No special technique is needed to communicate editorial norms that are open to complete reporting of all relevant studies. Publication pressure is an often-cited reason why authors have suppressed inconvenient findings. Therefore, editors, particularly editors-in-chief, are key to announcing these norms. It is a trickier call, however, whether and how to act against accumulations of evidence that invite suspicion through a "too good to be true" pattern of significant results. After all, it is possible that any single multi-study winning perfecta was obtained through complete and straightforward reporting -- just not very likely. Encouraging complete reporting, together with some of the other evidence reforms such as pre-registration and disclosure, can help. However, we should also avoid making *imperfection* a heuristic of trustworthiness! We would really be losing something if people felt they had to produce a non-significant study, to give a line of significant studies the veneer of credibility.

## Evidence Solution 3: Publish More Articles With Null Conclusions

### *Why?*

Accepting articles with mixtures of significant and non-significant studies does not necessarily challenge the publishing goal to produce articles that overall make a *positive* statistical point. However, many conclusions of interest are *negative*. The public and stakeholders want to know when there are no appreciable differences between genders or cultures, when a popularly held belief is untrue, or when a costly treatment or intervention has no effect big enough to justify investing in it. However, most articles published in psychology show conclusions that support, rather than question, the tested hypotheses.

A few observers over the years have remarked critically on psychology's bias toward positive conclusions (e.g., Coursol & Wagner, 1986; Fanelli, 2010; Greenwald, 1975). Now and then, some of them have wondered whether there should be an outlet for negative results, but without lasting success in establishing one (see Giner-Sorolla, 2012 for a historical review). A strange division of norms exists. Authorities such as the APA Publication Manual and the authors of writing guides (Bem, 2003; Sternberg & Sternberg, 2010) repeatedly counsel against concealing null results, as we have seen earlier. Yet bodies of published articles in psychology, as with the previous section's look at single articles, show an overreporting of significant results, even assuming their sample effect size is a perfect picture of the population effect size (Bakker et al., 2012; Francis, 2012). If we allow that some research programs might be testing an idea that is not true, and others might fail through error or misspecification, the so-called "file drawer" of research not reported because the main findings were not significant (Rosenthal, 1979) grows even further.

## *Uptake*

Despite many decades of observation and lamentation, the rate of positive conclusions in psychology articles continues to hover around 90% and upwards (Scheel et al., 2021). This steady state may lead hopes to focus on other formats that are more accepting of negative results, such as Registered Reports (Scheel et al., 2021), or preregistered replications (Kvarven et al., 2020). Positive result rates of 90%, if they really did represent all the research being done, would signal a discipline with very strong theories and few surprises: studying research questions that are 95% likely to be true, with 95% power. A more creative discipline, perhaps, would study questions that are likely to be true, but not done deals (suggesting anywhere from 50-80% prior truth likelihood) and which are strong enough to be detectable at 90% power within the resources of a well-funded laboratory. Under these assumptions, our positive results rate should be more like 45-70% for initial tests of reasonable ideas. Rates of conceptual replication should then be somewhat higher, depending on the ability of the initial methods to rule out false positives and to generalize to other settings.

Positivity bias can further be tackled by designing studies that provide a critical test between two opposed ideas, changing the spectrum from one bounded by "positive" and "negative" results to one bounded by two different "positives". This approach characterizes many classic research programs in social psychology (for reviews see e.g. Harmon-Jones & Mills, 2019 on challenges to dissonance; Kang & Swann, 2010 on self-enhancement vs. self-verification). But in the present era, it has been observed that, like toothbrushes, theories are mostly used and tested by their owners (Mischel, 2008). One recent exception is the movement to encourage adversarial collaboration involving people on both sides of a debate, crafting studies with outcomes that can distinguish between rival hypotheses (Clark et al., 2019; Ellemers et al., 2020; Kahneman, 2003).

Meta-analyses, too, can go beyond the bias inherent in published reports. They can grab effect sizes from unpublished research and secondary analyses, potentially presenting a more realistic picture. One recent study including 83 meta-analyses in psychology, unlike previous efforts, took care to

respect the homogeneity assumptions of bias detection methods, and concluded that publication bias in this sample exists, but had a relatively small effect on conclusions and effect size estimates (van Aert et al., 2019). Public archiving of nonsignificant results without formal peer review, as exemplified by the website psychfiledrawer.org, are another way to bring nonsignificant results into considerations of literature (Spellman, 2012). But we might ask why these results must enter through the back door in the first place, forcing meta-analysts themselves to do basic quality control that peer review is expected to carry out.

## *Costs And Limitations*

Interpreting a literature with an unbiased positivity rate of 45-70% might be difficult, because researchers do not trust reports of negative results. The reasons for this mistrust should be addressed ahead of time in planning and reporting research. When faced with a p-value upwards of .10, it is natural to wonder whether the result is a genuine negative. Maybe it can be explained by some error in procedure, a failure of the necessary conditions for the effect, or a lack of statistical power. After all, a null result under null-hypothesis statistics cannot rule out the existence of a sufficiently tiny-sized population effect tilted one way or another, no matter how large the sample and how accurate the measures. This problem comes from a failing of null-hypothesis statistics indicted by several authors, among them Szucs and Ioannidis (2017): "Researchers set H0 nearly always 'predicting' zero effect but do not quantitatively define H1" (paragraph 15). Because predicted effect sizes are not usually a part of psychological theorizing (Meehl, 1967), many fields of research don't have a good idea of what effect size is too small to matter, leaving some free to argue that this number can be very small indeed (Funder & Ozer, 2019; Götz et al., 2021).

Researchers who want other researchers to believe in the reality of their null results thus have a daunting task ahead of them. For this challenge, most of the costs fall on the researcher, not the evaluator. They must offer convincing evidence that their procedure worked. For example, they might report the results of pilot tests or manipulation checks. Authors who get positive results often feel they do not have to demonstrate their methods' validity (Ejelöv et al., 2020; Fiedler et al., 2021). Because researchers do not know in advance whether results will be null, this advice implies that they should improve their checking and testing across all studies. These improvements will be time-consuming but would have the beneficial effect of improving sensitivity to positive results as well (see "Challenges of Methodology" section, below.)

Researchers who publish negative results may also wish to offer statistical evidence in support of the null. The standard null-hypothesis framework is not equipped to deliver this information, despite the common error of interpreting nonsignificant results as evidence against the existence of an effect (Aczel et al., 2018). Fortunately, there are ways to assess evidence in favor of the null, also known as equivalence testing, in both Bayesian and null-hypothesis frameworks (Lakens et al., 2018; see Linde et al., 2021, for a simulation comparing three methods). Even null-hypothesis methods, though, need to be "Bayesian" to some extent, in that they require an effect size of minimal interest to be specified as a hypothetical limit on meaningful priors. While coming up with such a number is easy, justifying it may be more difficult, because heuristics for a consensus minimal effect size are currently not well developed (see Giner-Sorolla et al., 2024; Lakens et al., 2022). Perhaps, instead, the problem lies with the heuristic that the journal article package must deliver a definitive positive or negative story, instead of offering up valid results regardless of outcome, which might eventually be clarified by further research.

Finally, the null-results promoter must also convince someone that their result is interesting. "Bunk" must exist before it is debunked. That is, for a negative result to be interesting, the positive result would have to be plausible according to some lay belief, philosophical premise, or scientific theory. For example, if a scientist presents evidence that wearing a green t-shirt has no effect on people's ability to jump, nobody will care, because nobody ever believed that the color has performance-enhancing properties. But if they present evidence that wearing green does make people jump higher, maybe because it puts them in the mindset of a frog (just speculating here), that is a surprising result that people might be interested in. Granted, ideas like this lack the formal theory development that would let us anticipate the result rather than be surprised by it, but that is a different problem to be addressed later.

### *New Rules?*

Even keeping the rule that articles should tell a story with a definitive conclusion, that conclusion can be null if: a) the idea being tested has a solid basis in theory, or in a popular belief that deserves to be tested; b) the methodology used to test the idea can be shown to be valid, through independent checks and pilot testing; and c) the statistical analysis uses a method that can quantify and support results in favor of the null hypothesis. Doing so will help ensure that our literature is more credible. Dead ends can be identified and publicized, rather than stumbled into repeatedly by different explorers of science. As another benefit, meta-analytic aggregations of results would be based more solidly on peer-reviewed literature, rather than having to plumb a completely unreviewed file drawer. Of course, a more complete proposal would disseminate all valid findings on points of interest regardless of whether they are positive, negative, or inconclusive. This proposal is more feasible now that online publication allows storage and retrieval of research reports, unconstrained by the page limits of paper journals, but is complicated by the labor required to review each report and ensure it was carried out in a valid way, with all details reported.

## Evidence Solution 4: Report Independent Direct Replication Studies

### *Why?*

The multi-study solution to the problems of evidence unearthed in the 1970's ensured that strong articles would include some kind of replication of the initial effect. These came in the form of conceptual replications (using different measures, manipulations, or contexts), but also in the less-well-studied practice of "replication and extension," such as by adding a condition or moderator variable while keeping part of the design that replicates a previous study (Frank, 2015). While calls for reform in the early 2010's acknowledged the existence of replications by the same authors, they also found fault with a reliance upon them for evidence. After all, quite noticeably, the Bem (2011) article included eight conceptual replications, but still failed to convince most psychologists of the existence of precognition. As already discussed, "replication" to improve evidence has focused on direct replications by independent teams, copying the methods of the original to the greatest extent possible, rather than conceptual replications by the same laboratory (see Zwaan et al., 2018, and the replies in the same issue for a more thorough discussion of the issues around direct vs. conceptual replication).

Debates about direct replication have a long history in science. One common understanding holds that the ability of independent teams to reproduce a finding is a prerequisite to treating it as a scientific fact (e.g., Popper, 1959). However, Collins (1985) shows that even in fields generally seen as having strong theories and methods, such as physics, replication turns out to be controversial because of disagreements on what the essential elements of methods are. Likewise, a failed direct replication is not usually seen as conclusive evidence against a proposition in psychology, or at least, not sufficient to warrant the retraction of the original article. A quick look at the Retraction Watch Database (2018) of retracted articles confirms this. Psychology, in early 2024, shows only eight retractions classified as "results not reproducible," most of these having other reasons for retraction, such as investigation outcomes. By contrast, fields such as physics, materials science, and cancer biology show more retractions based on failed replications. Interestingly, it is not unusual for a retraction in other sciences to come as an admission of error from the researchers themselves, after they could not replicate their own findings in further studies. Self-retraction following errors unearthed by a failed internal replication has happened at least once in social psychology as well (the retracted article being Mahajan et al., 2011).

Looking over reactions to the replication issue, the problems start when we try to agree on a criterion for saying that one study replicates another with "success" or "failure." In algorithmic and simulation analysis of replication criteria, statistical methods comparing only two studies to each other suffer either from an excess of false positives or false negatives (Schauer & Hedges, 2021). Beyond this issue, there seem to be two divergent ways to approach the interpretation of replication findings. The more dramatic way is to see a failed replication as conclusive proof that the original study was at best erroneous and at worst an example of misconduct. This attitude is not uniquely a product of post-2011 Internet discourse; Schmidt (2009), in writing a review article about replication, lists the functions of direct replication as controlling for sampling error, artifact, fraud, and generalizing results. The possibility that replication might catch a false-positive finding thanks to commonly used selective reporting procedures, rather than outright fraud, is not mentioned in Schmidt's otherwise prescient review.

Against this backdrop, the heated rhetoric between non-replicators[3] and the non-replicated in the 2010's might have been expected (for a variety of perspectives on the tone of these discussions, see Bartlett, 2014; Bishop, 2014; Nicolas et al., 2019; and an excellent recent review and analysis by Derksen & Field, 2022). Without naming names, a recurring dynamic can be identified. Reactions to non-replication that explicitly cast the original research into suspicion, shading into serious accusations of fraud, provoke a counterswing: accusations that the non-replicators are incompetent, low-status (for a reason), sadistically motivated to tear things down, or at the very least have omitted important contexts or procedures from their attempts (as foreseen by Hendrick, 1990).

When judging unsuccessful replications, however, we should automatically privilege neither the original finding's positive result nor the replication's negative result. Rather, we should see the two as data points with a discrepancy to be resolved (Maxwell et al., 2015). Possible resolutions could hinge on: a) error of some kind in one of the two studies, verified through review of the procedure and data; b) the original positive result being produced through selective reporting, or through a statistical fluke, verified through the failure of further replication attempts (direct or otherwise), as in Bem (2011); c) methodologically but not theoretically meaningful differences between the studies, verified through detailed comparison of procedures, including the possibility that one study used more accurate or powerful methods than the other (Bressan, 2019; Gilbert et al., 2016); d) error in the replication finding, either through selective reporting of negative results or through a statistical fluke; or e) theoretically meaningful differences between methods and contexts used in the

replication and original. On the last point, further research or meta-analytic investigation might confirm the overall pattern of stronger effects under one context and weaker effects under another, encouraging further development of the boundaries and conditions of relevant theory (for some recent examples, see Luttrell et al., 2017; Noah et al., 2018; and the exchange between Landy & Goodwin, 2015, and Schnall et al., 2015).

There is a contradiction between two prominent critical positions toward replication. On one hand we hear there is no such thing as a "direct" replication, because culture, time, and other aspects of context make every replication new (e.g., Stroebe & Strack, 2014; Stroebe, 2019). On the other we hear that direct replications cannot advance theory because they offer nothing new (C. S. Crandall & J. W. Sherman, 2016). Instead, replications in social psychology that are meant to be direct sometimes turn out to involve meaningful variations. These variations establish generalizability if they succeed, and perhaps specify further boundary conditions if they fail. Fabrigar and Wegener (2016), in fact, define good replication practice as reproduction of the underlying psychometric variables in the new context, rather than literal imitation of the procedure originally used. In trying to follow such advice, researchers will be likely to turn up deficiencies in psychometric assumptions, as well as grapple with theories that are not well developed enough to specify how they should be tested in different contexts (Cesario, 2014; Haig, 2021), both issues I further cover in later sections.

There are at least two other reasons to publish direct replication studies. First, suppressing replication studies under the formerly reigning heuristic of novelty, gives the lie to scientists' claim that their method of knowledge is superior to others because it is self-correcting (Vazire & Holcomb, 2021). Under this view, publishing replication studies by external labs is akin to other strategies of scientific self-correction such as double-checking published statistics (Nuijten et al., 2020), reproducing analyses from data (Artner et al., 2021), or encouraging post-publication review of limitations and errors (Bastian, 2014). Although replication studies are not as straightforward to interpret as computational reproducibility is[4], they can still lead to beneficial error correction. Dismissing them makes the field look as if it has something to hide.

Replication also has a specific utility when dealing with a research literature that has been produced under conditions of selective reporting. In a sense, every replication happens with a pre-registration in the form of the original article's method and analyses, which it is trying to follow more or less closely. Current best practices go further and involve the literal pre-registration of replication attempts (Brandt et al. 2014). As we have seen, an original set of studies might have been collated from the most statistically significant findings among a series of conceptual replications, and further selectively analyzed for the best results in each study. But a replication chooses a fixed target a priori, re-establishing the conditions for interpreting inferential statistics that were abrogated in the original research. Putting together this rationale with the previous one, it stands to reason that external replication is less useful for verifying the results of studies that have been produced under a guarantee of no publication bias (e.g., as a Registered Report), than for verifying research that was more loosely regulated to begin with.

A larger point bearing on replication, which Devezer et al. (2021) have made, is that scientific reform procedures should be implemented only to the extent that they are themselves methodologically proven under a variety of conditions. This means that some procedures might be mutually somewhat redundant, as in our example of replication becoming less necessary when research takes on tighter reporting standards. Other times, reform procedures might be insufficient to establish the truth. Replication, for example, is not exempt from the risk of verifying conclusions

that are false. After all, unidentified confounds that threaten the validity of positive findings might also replicate very nicely.

## *Uptake*

After roughly a decade of renewed interest, what is the state of our journals' interest in publishing direct independent replication studies? Early indicators were not promising. Survey and archival studies of editors and journals showed pervasive anti-replication attitudes in psychology pre-2011 (e.g. Neuliep & R. Crandall, 1990). Martin and R. M. Clarke (2017) set out to look at submission guidelines in psychology to see if things had changed by 2015. However, very few journals in social psychology (4 out of 93) explicitly encouraged replication studies. Most were silent on the issue, or implicitly discouraged replications though submission criteria that privileged novelty. The picture was similar in other fields of psychology.

It seems, however, that social psychology journals have become more receptive to replication since 2015, due to turnover in editorial teams and changing priorities of sponsoring organizations. For example, a recent initiative of the APA to improve its journals' Transparency and Openness Promotion ratings (TOP; Nosek et al., 2015) ended up with all its journals stating openness to replication studies in their submission guidelines. Some, including all sections of the social psychology flagship *JPSP*, further specify that replication proposals will be peer-reviewed without reference to the eventual results (American Psychological Association, 2021). As of mid-2022, other well-known social psychology journals including *Personality and Social Psychology Bulletin, Journal of Experimental Social Psychology*, and *Social and Personality Psychological Science* have explicitly encouraged replication submissions.

But how are researchers responding to these invitations? I conducted Google Scholar searches in February 2024, specifying the four journals mentioned in the previous paragraph, the years 2018-2024, and requiring the word "replication" to appear in the title, with further scrutiny to cast out duplicates and ensure that the studies were in fact reports of empirical replication attempts. I assumed here that replication-focused articles would identify themselves as such in the title rather than risk the misinterpretation that they were claiming to advance a novel idea. Each journal published from 766 to 880 articles total in this period. Out of that number, *JESP* published 14 self-identified replication articles, *JPSP* published 10, *SPPS* 9, and *PSPB* only 4. Of these, *JESP* and *SPPS* had clear policies allowing replications over the whole six-year period, but that did not lead to an enormous flood of replications in absolute terms.

## *Costs And Limitations*

The open door for replication, but with few visitors, should alert us that psychology researchers are not finding the resources or motivation to conduct replications at the rate proposed by LeBel (2016): one replication study of other people's work for every four original studies. In fact, judging by publication rates, they are not even collectively producing one replication article for every *sixty* original articles.

As we have seen, there are many open questions about which and how many replications should be carried out. An interesting, if skeletal, triage framework to evaluate the costs and benefits of different replications in a situation of limited resources has recently been proposed by Isager et al.

(2023). Broadly put, the less certain existing findings are, and the more important they are to psychological theory, education, or application, the stronger the case for conducting the replication.

However, the ratio of replication to original studies required to make progress is also a point of legitimate debate (see Hendrick, 1990). We might ask the value of funding a replication study, compared to simply making sure the original study is high-powered and completely reported. In doing so, we should consider the benefit that direct replication series give, compared to single studies with the same total N, in testing the effect across heterogeneous contexts (Kenny & Judd, 2019; IntHout et al., 2015). Not every study, though, is realistically going to be replicated. To have the most influence on knowledge, policy should focus on replicating studies that turn out to be influential (for simulations in support of this last point, see Lewandowsky & Oberauer, 2020). We might, more generally, try to establish the optimal trade-off between researching novel ideas and confirming older ideas, both within and between research teams (Finkel et al., 2017a, 2017b; LeBel et al., 2017). Also, while knowing about a general norm of replication might encourage researchers to do science more slowly and carefully in the first place, simulations show that this incentive is not enough by itself to thwart the career benefits of low-effort approaches to research (Smaldino & McElreath, 2016).

On the individual level, fear of entering a public controversy might be a motivational barrier to running replications (LeBel, 2016). But it is also possible that researchers just prefer to build and inhabit their own edifice of knowledge, rather than inspecting the buildings of others. They may feel that checking other labs' ideas diverts time and money from their ability to test and promote their own. They may chafe at the additional, strongly recommended step of consulting with original authors and trying to recreate their procedures faithfully (Brandt et al., 2014), or at adding steps to ensure psychometric accuracy which, often, original authors omitted.

Until recently, it has been unknown what motivates researchers to choose to replicate specific effects (Giner-Sorolla et al., 2018). This mysteriousness sometimes leads to suspicion that replications are motivated by malice or bias[5]. A recent mixed-methods study of researchers' reasons for carrying out replications has found that while curiosity about a finding (69%) and recognition of its importance (50%) were important, doubt about a finding was also a prevalent motive (51%; Pittelkow et al., 2023). Indeed, if greater attention is given to failures to replicate than to successes, this might foster a view of replication as an extraordinary call to arms driven by doubt, not as an everyday independent test of a finding. Currently, relying on purely internal motivations to replicate other people's work hasn't been enough to produce a sizable body of published replications. If having more replications is desirable, other solutions need to be weighed, including the possibility of rewarding career specialization in replication (Romero, 2020).

## *New Rules?*

In the past, reviewers and editors might have felt uncomfortable handling replication manuscripts because there were not enough of them published to develop common knowledge for evaluating them. In a highly selective merit-based system, integrating replication reports requires special rules. Such rules might emphasize the responsibility of a journal or a field to publish replications of its own published articles in the spirit of self-correction (Srivastava, 2012; Lindsay, 2017). Readers are more likely to update their estimates of an effect on the basis of less successful replications that are published in the same journal as the original, and also endorsed by the original researchers (Eriksson & Simpson, 2013). Journals, too, might set standards for quality to reassure scientific

gatekeepers that they are applying similar high standards to replication research as original research.

Brandt et al. (2014) suggested a list of such quality standards, including well-justified statistical power, consultation with original authors where feasible, sensitivity to the impact of method changes, documenting and registering plans, absence or inconclusiveness of prior replication efforts, and selection of an influential study to replicate. In hindsight, perhaps the last two criteria are too dismissive of some other legitimate reasons to carry out a replication, such as an interest in the topic that leads a researcher to test the effect in their laboratory. But the first four criteria follow from an assumption of epistemological balance that favors neither the original nor the replication study automatically. They help to ensure that the conclusions of the replication are solid enough to begin to revise or reinforce the conclusions of the initial study.

Finally, evaluators of replication studies must not fall into the mirror image of the positivity bias, judging results that contradict the original study as being more publication-worthy than results that support the original study! As Hendrick (1990) observed, "... successful [direct] replications bolster confidence in old data but have little novelty value" (p. 47). This feature risks a particularly pernicious interaction with the traditional heuristic of novelty in publishing. P-hacking and selective reporting to obtain a *negative* result also demonstrably exist (for example, when needing to "prove" that a confounding factor's influence is nonsignificant; Chuard et al., 2019). Worse yet, many errors in procedure can be quietly tolerated if a researcher feels they need to get negative results to be published. Outcome bias can most effectively be avoided by requiring results-independent peer review of the replication's rationale and methods, either prior to the replication's results (e.g., Registered Reports), or without seeing them (e.g., results-blinded review).

## Evidence Solution 5: Increase Statistical Power

### *Why?*

Statistical power, a metric developed by Jacob Cohen (e.g., Cohen, 1969, 1992), is the probability that a study sampling a population containing an effect of a certain size will return a significant result under null-hypothesis testing. Power increases as the number of participants in the sample increases, but also partly depends on the study's design (via effect size) and the criterion of significance being used. To a reader familiar with statistics but not with developments in social psychology, it may seem strange to hear that low statistical power lessens the credibility of a study's results. After all, isn't power analysis only useful before the study is run, to determine the sample size (as argued by, e.g., Levine and Ensom, 2001, Senn, 2002, Wilkinson Task Force, 1999)? And given that the problem seems to involve too many significant results being reported, how on earth can increasing the power to produce more of these significant results possibly help?

Although the first of these objections is technical, and the second naive, they can be answered in two ways. First, the lower the power of studies, the more a literature that selectively reports only significant results exaggerates the size of an effect. That is, simulations tell us that putting 1000 participants into 25 underpowered forty-person experiments, when there is no effect in the population, is more likely to produce publishable false positives, compared to putting them into two 500-person experiments (Stanley et al, 2022; Sterne et al, 2000). Chasing a small effect with small studies under positively biased publication norms is a false economy, because of the large number

of null-result studies, even if these can be explained away as failed method variants or unsuccessful pilot studies (LeBel & Peters, 2012).

The telltale signs of an exaggerated literature are low-powered studies, large effect sizes (necessarily, because the studies are not powered to consistently detect smaller ones), and significant p-values that on average are closer to .05 than would be expected if there were a true effect with the observed size in the population. This latter feature comes about because significant p-values under the null hypothesis come from a uniform distribution and are just as likely to be .049 as .001. However, significant p-values under the alternative hypothesis come from a right-skewed distribution and are much more likely to be .001 than .049. In general, significant effects obtained under low power are less likely to form a consistent, replicable literature than those obtained under high power (Maxwell, 2004; Mayo, 2018).

Second, formal statistical analysis also can clarify why and when low-powered significant results are less likely to reflect an underlying effect. Understanding these analyses starts with acknowledging a common erroneous belief: that a p-value is the chance that the effect observed in the sample is not true in the population (S. F. Anderson, 2020). This belief, commonly held even among researchers and educators (Badenes-Ribera et al., 2016; Cassidy et al, 2019), is a form of the *affirming the consequent* fallacy. If we see that all dogs have fur, and like a two-year-old child learning their words, mistakenly conclude that everything with fur is a dog, that is affirming the consequent. Formally stated, the incorrect belief is that p(fur|dog) = p(dog|fur).

The p-value, then, is formally stated as p(positive|null) -- that is, the chance that sampling from a null-hypothesis population would give you the kind of value you are observing. What people wrongly think the p-value is – the chance that the observed positive result is wrong, p(null|positive) -- can be calculated. It is known as the false discovery rate (FDR; Ioannidis, 2005) or false positive risk (FPR; Colquhoun, 2019). However, this calculation depends on two additional pieces of information that social psychologists often find difficult to establish conclusively. First, it needs to know the power of the statistical test, which depends in turn on knowing the population effect size; second, it needs to know the prior probability that the null hypothesis is true. Nonetheless, arbitrary values of power and priors can be entered into a model, and playing around with an interactive simulation is a good way to learn how FPR works (e.g., Zehetleitner & Schönbrodt, 2022). S. F. Anderson (2020) also presents simulations that underscore an important point: a study's sample size does influence the FPR, but so does the prior likelihood of the hypothesis.

For example, assume 20 participants in each of two conditions, a population effect size $\delta = .3$, and a hypothesis with a null prior of .5 (i.e., the alternative hypothesis $H_1$ is 50% likely to be true). Under these conditions, $p < .05$ corresponds to a false positive rate of .25. Tripling the n per cell, to 60, lowers the FPR to .12. However, choosing a more plausible set of hypotheses with a lower null prior of .2 (that is, $H_1$ is 80% likely to be true) lowers the FPR even further while keeping n = 20, to .08. Finally, an FPR comparable to the nominal significance level of .05 can only be approximated by combining a high n of 60 with a highly probable prior, setting $H_1$ at 80%, thus yielding an FPR of .03.

Wegener et al. (2022) further bolster the case that power might be less important than choosing a plausible effect to study in the first place. They show through simulations that power's impact on false-positive rates begins to level off after 50% power, and when the question under study has prior odds higher than 50%. We will return to these considerations when looking at optimal decision rules about studies with lower and higher statistical power. But to conclude, there is no reason to believe *more* in the truth of a significant effect because it came from a low-powered study

-- the fallacy of a "heroic effect" dissected by Loken and Gelman (2017) -- and in fact there are some reasons to believe it *less*.

## *Uptake*

The statistical power of typical studies in psychology has come under intermittent scrutiny ever since Cohen (1962). Analyzing the power of published studies in 1960 in the most prestigious social psychology journal of that time (Journal of Abnormal and Social Psychology, the precursor to *JPSP*), he found that most of them only met his 80%+ power criterion to detect an effect that was "large" by his standards. For Cohen at that time, low power was a problem because it suggested that many studies were obtaining inconclusive, nonsignificant results, a waste of research effort in a publication world biased toward reporting significant results. He further believed that his "large" criterion denoted effects which were "so obvious as to virtually render a statistical test superfluous" (p. 146). He concluded that most studies did not have the power to consistently detect more reasonable targets of research, so many nonsignificant studies must be languishing unreported in the file drawer.

Further investigations by Sedlmeier and Gigerenzer (1989) and Rossi (1990) showed that, approaching the 30th anniversary of Cohen's study, power in comparable journals had not improved at all. The next major overview of published research power relevant to social psychology comes from Fraley and Vazire (2014), covering the years 2006-2010. In this analysis, where effect sizes were based on *r* rather than Cohen's *d* as a common metric, social psychology journals had from 78-84% average power to detect $r = .3$ but only 43-49% power to detect $r = .2$ (personality psychology journals did a bit better on average). Fraley and Vazire argued that $r = .2$ was a realistic benchmark for effect sizes based on meta-analyses of social psychology effects, but Cohen believed $r = .3$ to be "medium", even though mathematically his "medium" score $d = .5$ corresponds instead to $r = .25$. However, holding mathematical effect sizes constant, the 2014 overview can be said to show some improvement in the average power of research compared to previous overviews. This may be due to developments over 50 years facilitating larger studies through undergraduate participation pools (e.g., establishment and electronic management of pools, computer printing and photocopying, increased student numbers).

In the early 2010's, ongoing concerns about the power of social psychology research found another apparently fortunate answer, with increasing use of large-scale online crowdsourcing pools such as Amazon Mechanical Turk and Prolific Academic (Behrend et al., 2011; Buhrmester et al., 2013; Peer et al., 2017). Two analyses so far have focused on changes in sample sizes in social psychology journals during this period.

Sassenberg and Ditrich (2019) looked at 2009 and 2011 compared to 2016 and 2018, and found an increase from mean outlier-corrected sample sizes of 120 and 117 to 179 and 195 respectively. Although this increase was not further analyzed in terms of power, it was accompanied by a rise in the use of online samples, from 6% of all samples in 2009 to nearly 50% in 2018 (see also C. A. Anderson et al., 2019 for further evidence of increasing reliance on online samples). Motyl et al. (2017), as we have seen in our look at effect size reporting, compared 2003-4 with 2013-14. Although their statistics are hard to compare with Sassenberg and Ditrich's because they handled the data differently, they also found a modest, statistically significant increase in sample size and a corresponding increase in power, such that 25% of studies in the later time period, compared to 15% in the earlier, had 80% or more power to detect $r = .21$. Both figures are notably lower than Fraley and Vazire's 43-49% for $r = .20$ in 2006-2010, possibly because those authors used more direct and

design-sensitive methods to quantify power. Overall, it seems that technological advances are being used increasingly over time to help allay, if not completely satisfy, concerns about inferences from relatively low sample sizes in social psychology.

## *Costs And Limitations*

The most obvious cost of increasing power is the cost of recruiting more participants. Competitively, a laboratory that runs studies with twice as many participants will have half as many publications. This drawback can be offset somewhat by avoiding the wasted effort of running low-powered studies, and by editorial standards that prefer papers with higher power (if these standards are not in place, simulations show that the low-power game is a winning game: Bakker et al, 2012; Smaldino & McElreath, 2016). A healthier academic career evaluation system, in which quality of evidence is weighted more heavily than raw numbers of papers, would also make the cost of high-powered research less troublesome.

Still, other concerns about the downstream consequences of increased costs have surfaced in response to calls for higher sample sizes. Perhaps increased requirements would exacerbate inequalities between institutions with different levels of research funding. Perhaps they would disadvantage research that uses highly involving methods, or that studies demographically less prevalent and harder-to-reach populations. Perhaps they would privilege research on relatively homogenous populations, or privilege methods that are easy to administer to large numbers of people over more rigorous tests of behavior or cognitive processes (C. A. Anderson et al., 2019; Baumeister, Vohs, & Funder, 2007; Doliński, 2018). To put it succinctly, emphasis on the internal validity afforded by high sample sizes should not lead us to abandon questions of external validity (cf. Sue, 1999).

These concerns cannot be dismissed easily. Increases in sample sizes and online sampling are accompanied by increases in reliance on self-report (vs. behavioral or implicit) measures (Sassenberg & Ditrich, 2019). To anticipate a later section, there is also evidence that social psychology samples have a diversity problem. In terms of ethnic representation within diverse societies, Roberts et al., (2020) point out that even among published social psychology research with race as its topic, close to 75% of participants are White. This may reflect a focus on studying dominant-group prejudice more so than on studying identity, coping, and other facets of racialized people's experiences. More broadly, the overwhelming preponderance of participants in psychology studies come from majority-White, English-speaking and Western European societies (Hruschka et al., 2018). Convenient online samples drawn from these societies may help solve statistical power worries, but can these findings generalize to other global settings?

One answer to these concerns is to realize that the power formula has other terms than sample size. Power can be increased through larger effect sizes: using improved methods with higher reliability and stronger manipulations, using within-participants designs to equalize variance across individuals (Smith & Little, 2018), using pretesting and covariates to account for theoretically uninteresting variation in outcomes. In fact, these solutions have always been common in cognitive psychology and social cognition.

We can also recognize that desired levels of power depend crucially on the size of effects we are looking for. Small samples or difficult methods can best support a research focus on strong effects, which require lower sample sizes to detect with adequate power. Difficulties in studying populations often reflect a lack of access due to economic or social discrimination. Therefore, it may

be more responsible for researchers to identify the strongest psychological influences on an issue for these populations, rather than chasing a single, subtle, decontextualized influence for theoretical reasons. But, critically, findings from the field should have some way to influence ideas about basic processes. This approach, in turn, would require a reconsideration of what counts as legitimate and prestigious research (Lewis, 2021), reversing the tilt away from applied research over the past twelve decades of psychology (Giner-Sorolla, 2019).

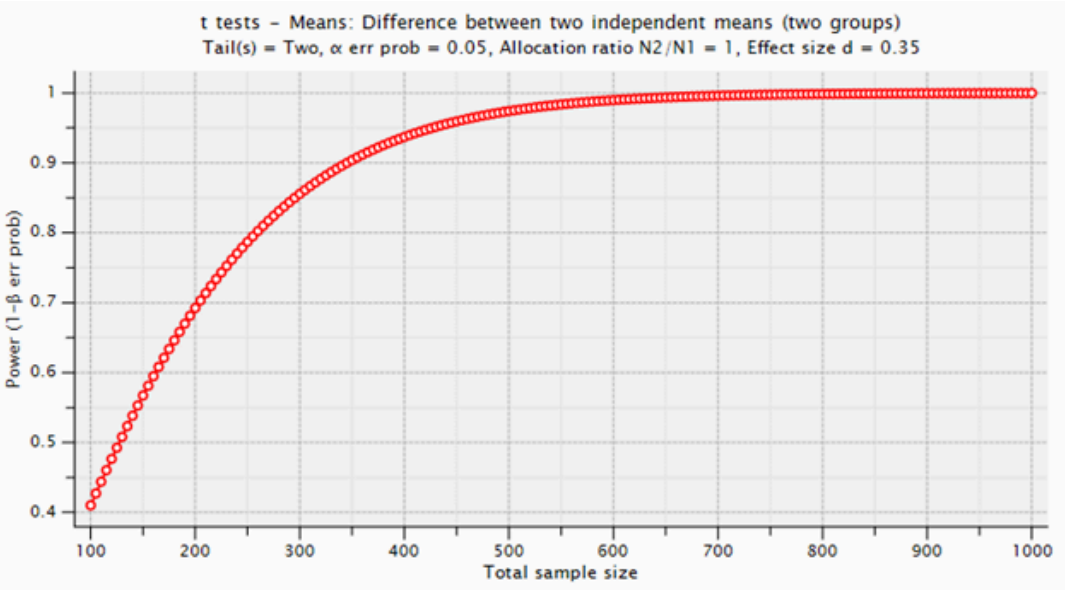
Another perspective on small samples comes from Lange (2020), who considers how replicability can be achieved when doing research on small populations with neurological conditions. Studying only one small sample is likely to yield a false-negative, but if multiple small-sample studies are filtered through positive publication bias, the answer may end up a false-positive. The answer is that any effect size, at any significance level, can represent legitimate data which should be aggregated into a larger picture. In turn, this approach requires changes to the publication scene. It requires the sharing and discovery of data sets regardless of their individual conclusions, using forums beyond the traditional peer-reviewed article, such as open databases and working papers. The distributed approach also requires multi-laboratory teams collaborating intentionally, challenging the heroic, lone-wolf model by which labs and careers are often evaluated. More relevant to social psychology, one promising initiative that seeks to test selected questions through multi-site research across cultures is the Psychological Science Accelerator (Moshontz et al., 2018). To date, this organization has facilitated international large-scale research on such topics as social dimensions of face perception (Jones et al., 2021) and social distancing motivation during COVID-19 (Legate et al., 2022).

## *New Rules?*

Despite its central place in research methodology, power analysis in social psychology currently lacks agreement on two parameters that would give it a strong basis for consensus. From Jacob Cohen we have inherited a suggested power criterion of .80 and some suggestions about "small," "medium," and "large" effect sizes in psychological research. But by Cohen's own admission, none of these values was meant to be graven in stone (Cohen, 1988, p. 12, p. 56). Target power levels and effect sizes were arrived at through impressions and reasoning processes that can and should be questioned.

The easier number to think about is the .80 criterion for power itself. While it is better to have more than less power, increasing levels of power come at a cost. Three levels of power are typically proposed: .80, Cohen's original; .95, which is high but gives symmetry between false negatives given H1 (beta = .05), and false positives given H0 (alpha = .05); and the compromise number of .90. To run studies intentionally expecting more than 20% false negatives seems wasteful (Freedman et al., 2001); but the cost of symmetry at .95 power is also high. The claim that .80 power represents a special inflection point (Cumming, 2012) has been contested mathematically (Bacchetti et al. 2010), but it is also clear that returns diminish progressively as sample size is increased. Those who are interested can use the graphing function of the GPower software (Faul et al., 2019) to look at power on the y-axis and sample size on the x-axis, holding effect size and alpha constant. Figure 1 plots power to detect d = .35 in a t-test sampling between 100 and 1000 participants total. In this graphic the curve continually gets less steep, with the angle of the slope reaching 45 degrees somewhere around .80 power, and 27 degrees around .90. To get from .70 to .80 power only requires adding 50 participants, while to get from .80 to .90 requires adding over 100, and from .90 to .99 requires about 250 more.

*Figure 1*: Plot of Power Against Total Sample Size to Detect D = 0.35 in a Between-Participants T-Test




*Note: Generated with G*Power software version 3.1.9.6 (Faul et al., 2019).*

Power levels might also be chosen based on trade-offs between resources and the cost of a false negative. For example, the Replication Project: Psychology (Open Science Collaboration, 2015) set a standard minimum power level of .90 rather than Cohen's .80, presumably to support the interpretation that negative replication results were not simply due to low power. However, a minimum power of .80 for original research, in which the costs of a false negative are borne by the researcher, sounds reasonable. This minimum can be raised in situations where a false negative has identifiable drawbacks, such as a replication that might call into question a well-established finding, or a costly procedure where skimping on participants could endanger the whole rationale for spending on the study.

Contrary to some concerns, though, if the resources are there to use, and there are no trade-offs with other projects, no level of power is "too high." Worries about the overinterpretation of small, significant effects under high statistical power are only warranted if we assume old-school statistical thinking -- that is, taking p-values as the only indicator of the magnitude of an effect. In scientific reporting, as we have seen, effect sizes are becoming more and more visible. There might, however, be a problem communicating small effects to the public. People may expect that, when scientists report a difference between groups or a relationship between two variables, there is a larger effect size than actually exists (Hanel et al., 2017). But this problem can be solved by mentioning the size of the effect in lay terms and raw units. For example, a small effect in a large-sample social media study might be described as "For every extremist website a person followed, one more hateful word appeared per every 44 posts." In short, press-release worries should not constrain researchers from using all available resources within reason to improve their study's power.

The harder number to establish for power analysis, however, is the population **effect size.** This figure is crucial for making any kind of decision or evaluation using power analysis. In fact, power criteria are meaningless without establishing effect size first. A study that has 90% power assuming a large effect size can also be said to be poorly powered assuming a small enough effect size. But, by definition, the population effect size is unknown. It is no wonder that 31% of social and personality psychology researchers surveyed by Washburn et al. (2018) cited the uncertainty of effect sizes as a

reason to avoid using power analysis. Most of the common methods to estimate effect sizes in social psychology fall apart on further scrutiny, and no clear consensus has emerged on a heuristic to use.

Ideally, effect sizes in power analysis should be estimated using **theory**, much as physics theories can quantify the speed of a particle and the uncertainty surrounding that speed before observations are made. But, due to the inevitably complex nature of social behavior, theories in social psychology are not set up to do that (Meehl, 1967). Rather, they express effects only in directional terms, anticipating a positive or negative relationship between variables, or a more complex pattern that can be broken down into directional relationships. In principle, there is nothing stopping theorists from initially proposing an effect size together with a direction. Larger effect sizes might be proposed for theories that specify powerful and central rather than subtle and peripheral mechanisms, for example. Estimates of effect sizes in any given study might be further adjusted for the strength of measurement and manipulation of the variables involved. Still, given the state of our theorizing, the decision to state effect sizes a priori in a theory would be arbitrary and subject to revision as empirical results arrive.

Another frequent approach to estimating effect sizes is to use previously reported empirical results as a benchmark. This approach can be broad, as when the average sizes of many meta-analyses in social and personality psychology provide a benchmark for any topic in the field (e.g., Fraley & Vazire, 2014, drawing on Richard et al., 2003, derive a field-wide benchmark of $r = .21$). Benchmarking can become micro-specific, as when a replication study uses the reported effect size of the original study to guide its own sampling, or when researchers use a pilot study to gauge a novel effect in planning a follow-up. Benchmarks can also be drawn from a middle level of specificity, aggregating multiple studies that have a topic or methodology in common. Benchmarks based on meta-analyses generally find that the median overall effect size in social-personality psychology is lower than Cohen's "medium" value of .30 (Bosco et al., 2015; Gignac & Szodorai, 2016; Lovakov & Agadullina, 2021), with results closer to .20 (Funder & Ozer, 2019), in line with the Richard et al. (2003) estimates.

Two problems, though, arise with benchmarks. One is heterogeneity. Richard et al. (2003) based their estimate on a set of effect sizes from meta-analyses on different topics, which ranged from close to $|r| = .00$ to $|r| > .70$. Some of these effects were not necessarily theoretically predicted to exist in one direction or another (e.g., gender differences) while others were based on very strong and well-established effects (e.g., greater persuasiveness of experts). Even when researchers use a narrower topic range to estimate effect sizes, we can question how meaningful these averages can be across a variety of paradigms and methodologies -- factors that play a critical part in effect size. For instance, a researcher may not be content with the average effect size in a meta-analysis, but may want to adopt methods and contexts which yield strong effects, as long as they avoid confounds and other validity problems. While meta-analytic benchmarking returns an estimate that is plausible for an as-yet-unknown effect, it is less convincing as a way to obtain an *optimal* estimate.

Positive reporting bias also bedevils the source material for benchmarking. Procedures have been developed in meta-analysis for detecting and correcting publication bias (McShane et al., 2016). However, the best procedure is still a topic of development and controversy, hinging on the ability of data to meet assumptions (McShane et al., 2016; van Aert et al., 2019). When basing effect size estimates on single published studies (e.g., in replication), there have also been methods proposed to correct estimates for positive effects selection (S. F. Anderson et al., 2017; Perugini et al., 2014; Simonsohn et al., 2014). Pilot studies from one's own laboratory are also not immune from selection

bias. The fact that you have chosen to follow up initial significant findings, instead of abandoning them due to a disappointing, non-significant initial result, is a form of selectivity that needs to be accounted for (Albers & Lakens, 2018).

A different approach than all of the above ones would base power estimates, not on the typical effect size expected, but on the minimum effect size considered meaningful, or smallest effect size of interest (SESOI). This approach ensures that the study will have adequate power for measuring any reasonable directional effect in the field. However, as emphasized in the discussion of null results, *any* effect size can potentially be "of interest" to science, including those which can be shown to be almost indistinguishable from zero. The question can be rephrased: is there a range of effects for which, even if statistically significant, we would say that the effect is of no practical consequence? That is, are there effect sizes small enough to make us hesitate to announce that our science has shown a directional effect, for fear of being misunderstood? The effect size just outside of that range, then, is the SESOI. When put this way, however, it seems that the better option would be to show interest in all effect sizes, but with appropriate communication and cautious interpretation of the result's small size.

If any size of finding is of scientific interest, perhaps an alternative basis for power analysis might take as a benchmark the minimal *practical* interest of a finding we are seeking to verify (Giner-Sorolla et al., 2024). The debate over the value of effect sizes that Cohen would have categorized as "small" has gone on intermittently for decades, with some arguing for the value of pursuing such effects (Abelson, 1985; Funder & Ozer, 2019; Gotz et al., 2022; Prentice & Miller, 1992; Rosenthal, 1990) and others more skeptical (Anvari et al., 2021; Ferguson, 2009). A common argument is that small effect sizes found in one context, such as a single trial in the psychology laboratory, might accumulate or intensify in other contexts, such as through repeated exposure in a variety of real-life settings. However, this is akin to saying that a dollar is a lot of money because it could have won the Lotto. It is not guaranteed that cumulative effects will cumulate, or that effects in the laboratory will generalize to the field (Anvari et al., 2021). To establish those assumptions as facts in any one context, science needs to fund and value long-term research and field research to detect the larger effects said to follow from smaller ones.

Cohen's early attempts to estimate categories of effect size appealed to subjective notions of just-noticeable and blatantly-obvious differences, such as differences in height between adolescents of different ages. At times his assertions are debatable: "[a]n 8-point mean IQ difference is large enough to be noticeable; this is the order of magnitude of the difference between people in professional and managerial occupations" (Cohen, 1962, p. 147). However, a few more empirically sound efforts have followed. For example, recent studies have established the minimum noticeable amount of objectively measured mood change in participants (Anvari and Lakens, 2021). These efforts provide valuable information, but have one important shortcoming. Between Cohen's time and now, social psychology has developed an interest in unconscious processes, and an acknowledgement that introspective processes are not sufficient to explain all of human behavior. Naked-eye classifications of effect sizes may tell us something about which effects are accessible to consciousness. However, limiting investigation to phenomena so large that the average person can sense them puts limits on psychological research that we may not want to accept.

Ultimately, as with the power criterion of .80, effect size for power analysis is best thought of as a cost tradeoff. For applied research, stakeholders can express, in unstandardized raw units, the kind of effects that they would consider worthwhile to implement or act on, and researchers can power their experiments to that standard. Better yet, a formal cost-benefit analysis of the intervention

could pinpoint a range of sizes that are of interest depending on the price tag. A very inexpensive intervention, such as rewording the hand hygiene messaging in a workplace to invoke social norms, might only need to deliver a small benefit to be worthwhile, while a costly and effortful one, like an immersive diversity training program, would have to show large outcomes to be worth the cost.

Basic researchers should not see themselves as immune to economic considerations. Theory-testing studies also cost time, money, and other resources. Their benefits are measured not in the outcomes of interventions, but in their ability to inspire further expansion of human knowledge. So what if the McDuck Institute conducts an experiment on 100,000 participants in its 1000-cubicle, 500-employee mega laboratory, and finds a tiny-sized but significant effect at 90% power? If other labs cannot even hope to get 10% power to find that effect, it is not going to be replicated conclusively. Nor is the effect going to inspire research extending and building on it. Perhaps it is more justifiable to spend those resources running 100 studies that can only establish larger effects -- as long as the results of those 100 studies are fully reported, regardless of outcome. The resources involved in each study can then involve a greater segment of the research community. In such cases, field-wide discussion could recommend the kind of effect sizes that are most productive to pursue (Giner-Sorolla et al., 2024). If large samples can be gathered economically, they should be, but with the understanding that some effects they find to be non-zero may still be too small to continue pursuing through other means.

Lakens (2022) likewise recognizes resource limitations as a justifiable basis for sample size decisions. These limitations should be reported transparently, without tinkering with effect sizes to retroactively justify a pseudo-theoretical idea that fits the resource level. One form of power analysis, effect-size sensitivity analysis, can give as an output the minimum effect size that a given number of participants recruited with limited resources can attain, fixing a target power level such as .90. Discussion can then move on to whether that effect size is too large to be a realistic outcome of the research.

## Evidence Solution 6: Require Full Disclosure

### *Why?*

Researchers who admit engaging in selective reporting of results often cite the perfectionistic standards of competitive journals, as exemplified by editorial requests for a simpler story (Motyl et al, 2017; Washburn et al., 2018). One answer would be to reinforce a public norm against selective reporting, so that all authors would know that the same rules apply to everyone. We have already seen norms against omitting information printed in the APA Manual and elsewhere going back decades. Merely having them in the background was evidently not enough circa 2011. Since then, more overt initiatives have been proposed to ensure that Methods and Results sections in research reports stand as complete records of the research as actually conducted. Some of these initiatives involve disclosure through *showing* (posting open materials, open data, and open code) and others involve disclosure through *telling* (adding statements that vouch for the completeness of reporting).

The “showing” set of solutions has been greatly facilitated by the development of public online repositories such as the Open Science Framework which offer free, searchable archiving. Posting of data and materials has been encouraged in several ways. There are opt-in recognition schemes, most notably the Open Science badges granted to qualifying manuscripts in journals such as

*Psychological Science* (Kidwell et al., 2016). But increasingly, disclosure is encouraged through the kind of opt-out requirements that have been shown to be more effective than opt-ins for encouraging disclosure in other areas of human activity (Ioannou et al., 2021). In opt-out schemes, openly posted data and materials are required for published articles, but with the possibility of exceptions if there are strong ethical or legal reasons to keep the data confidential, such as sensitive personal information or proprietary data.

Posting open materials helps other researchers replicate and follow up on the findings and allows a full look at the study in peer review. Important details that emerge from looking at the full materials may not be discoverable through a Methods section write-up, especially when fitted to a tight word limit (Ledgerwood & Sherman, 2012). Posting open data also allows checking results in peer review, as well as testing alternative analyses and hypotheses that might modify the authors' conclusions. To this end, openly posting the code used to analyze the data can also help. Posting code is currently not often required in social psychology journals but is sometimes done voluntarily. Finally, open data helps the meta-analyst who needs a specific angle on the data, such as gender effects, that the authors' own Results section does not provide.

Open data is a different approach than the data sharing procedure that has been ordained by the APA's ethical guidelines and publication manual since at least the third edition in 1983. Under this framework, researchers should keep their original data and files for a minimum of 5 years after publishing in an APA journal. They are expected to share these data with competent professionals with a legitimate reason for access. Moreover, if there are costs of sharing data, the requester is expected to bear it. However, studies of data-on-demand requests from authors in APA journals have shown a low, if improving, rate of compliance. For example, Wicherts et al., 2006 got data from only 25% of APA-journal-based requests in 2005; Vanpaemel et al. (2015) got 38% of requested data sets in 2012; and Tedersoo et al., 2021, got somewhat better rates in the low 60% range in 2020 from psychology papers in *Science* and *Nature*.

These studies make clear that there are few consequences for non-sharing researchers. In some cases, researchers have barred requesters from legitimate uses (e.g., insisting that the data be used only to reproduce reported analyses rather than new analyses that might give a different perspective on the claim of the original research), or imposed onerous fees beyond reasonable costs of preparation for sharing data sets (Oransky & Marcus, 2016). Although the latest edition of the APA Publication Manual (7th edition, 2020) has expanded on the procedures and norms of data sharing to protect somewhat against such abuses, the on-demand system seems to err on the side of protecting researchers from bad-faith use of their data, rather than on the side of sharing and verification.

Moving from open data (showing) to disclosure statements (telling), the most widely used signal of full disclosure involves a 21-word standard statement to be included in research manuscripts (Simmons et al., 2012): "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study." The idea behind such disclosures is that researchers want to be honest but need confidence that other researchers are also following a norm of complete reporting. The clauses refer to undisclosed practices identified by Simmons et al. (2011) as p-hacking: running participants in waves until a significant effect emerges, choosing the most favorable result out of a variety of outcome measures, and selectively applying rules for excluding participants to attain the desired conclusion.

Importantly, disclosures are not meant to involve full analysis and reporting of every aspect of the research. Rather, they require all research elements to be mentioned, and this can be done in

footnotes, appendices, or supplementary materials. What's important is to allow careful readers to evaluate the significance tests against a background of how many chances were taken in total.

"Showing" disclosure through open data and materials might seem to reach the same goal more concretely than these "telling" statements. However, showing and telling complement each other. Disclosure statements cover procedures such as participant selection that might not be evident in the static materials. They also are a formal affidavit that the shown materials and data are themselves a complete account.

## *Uptake*

Outstripping the slow apparent improvement in responses to traditional data-sharing requests, more and more journals in psychology have led the way with disclosure and sharing requirements. Part of the trend can be attributed to the Transparency and Openness Promotion (TOP) initiative, which has encouraged adoption of these principles through direct contact with sponsoring societies and journal editors (Nosek et al., 2015). As of February 2024, looking only at social psychology journals without further topic specialization, the *Journal of Personality and Social Psychology*, *Journal of Experimental Social Psychology*, and *Personality and Social Psychology Bulletin* require disclosure statements in the manuscript, while *Social and Personality Psychological Science* requires them as part of the submission process, and *Social Psychology* encourages them.

Even more of these journals require data and material links to be made available with submission, allowing for opt-outs which must be explained and justified: the first four mentioned, plus *Asian Journal of Social Psychology, Social Psychology, British Journal of Social Psychology, Journal of Social Psychology*, *International Journal of Social Psychology,* and *International Review of Social Psychology* (the *European Journal of Social Psychology* requires data but not materials). Among generalist social psychology journals, only *Basic and Applied Social Psychology* does not require data or materials, only encourages them.

There is evidence that data and materials posting have increased over the past decade from a lower baseline. Rochios and Richmond (2022) report that in the general psychology journal *Psychological Science*, the degree and utility of data and materials sharing in social psychology articles increased across the board from 2014-15 to 2019-20. Hardwicke et al. (2022) also found low rates of data and materials sharing in a sample of psychology journals from 2014-2017.

Despite upward trends, there are still questions about the usefulness of data made available and the enforcement of journal policies. Towse et al. (2021) found that while the inclusion of open data sets in psychology journals increased from 2012 to 2017, of these data sets, 51% were not complete, or otherwise inaccessible. Rochios and Richmond (2022) found similar failures in their *Psychological Science* samples, but to a lesser degree. Their coding scheme found a non-zero range of scores indicating that data provided did not fully match the analyses, but the prevalence of such scores seems to have decreased over the six-year period they surveyed. Even if available, open data does not guarantee that analyses based on it will exactly reproduce the reported results, as shown by Crüwell et al. (2023) and Hardwicke et al. (2021).

## *Costs And Limitations*

Disclosure statements may meet resistance from authors because they are seen to protest too much, like a campaign advertisement that proclaims "Bob Talbert is not linked to the Mafia" but, oddly, does little to dispel doubts about candidate Talbert's integrity (Wegner et al., 1981). All the same, the bad reputation that has accrued to psychology in general, and social psychology in particular, might call for exceptional public reassurance. Possibly at some time in the future, if the disclosure norm becomes as established in everyday ethics as norms against plagiarism or data fabrication, disclosure statements will wither away. At most they might be absorbed into the publisher's online list of checkboxes affirming ethical compliance, as they currently are at *SPPS*. But while disclosure statements are not affirmed in all journals, they still serve a purpose among the journals that require them.

There are trickier issues in requiring open data. The anonymity of research participants is usually guaranteed at the point of informed consent, and special exemptions may have to be made for sharing de-identified data. Even in the absence of explicit identifying information, a particular combination of demographic traits may render an individual identifiable. Good advice for handling problems of anonymity and confidentiality in open-data research can be found in Alter and Gonzalez (2018) and other authors in that special section of the *American Psychologist*. It is encouraging that all journals in social psychology currently requiring open data allow for exceptions -- for example, if an analysis depends critically on personal data that may prove identifying. The challenge is for editors to make sure they are used for legitimate reasons and not as an excuse.

Another exception to open access may be claimed for proprietary data that researchers are legally bound not to disclose. However, if contracts also prohibit the data from being accessed for integrity monitoring by authorized institutions, the data exist in a black box beyond any checking. Yes, data from private sources are often highly relevant and interesting. But journals should consider the ethical jeopardy that arises when these sources must be taken completely on faith. For example, the open-access consortium of PLOS journals has disallowed articles using proprietary data for the better part of a decade (Bloom et al., 2014). Currently, at the general-interest PLOS ONE, this policy is worded to require third-party data sets to be available upon request to other researchers "in the same manner as the authors" (PLOS ONE, 2019) with appropriate exceptions to protect the confidentiality of individuals.

The uses to which others may put fully open data is also a concern that surveys have consistently identified as a barrier to uptake (Abele-Brehm et al., 2019; Fecher et al., 2015; Houtkoop et al., 2018; Washburn et al., 2018). One concern is that secondary use will not credit the original researchers, especially if collecting the data or generating the materials was costly or effortful. Indeed, requiring scientific resources to be "open" and "free" does not always align with justice (Johnson, 2014), so concerns about exploitation need to be addressed. The Mertonian norm of "communism" (in the sense of communalism; Merton, 1942) among scientists may not be fit for a non-utopian world where resources, opportunities, and credit are not themselves shared communally. Although the case has long been made that the data from publicly funded grants should be made available to the funding public (Ceci & Walker, 1983), the obligations attaching to private or institutional funding, or extending across state and national boundaries beyond the original funder, are less clear.

One solution to this concern is to require that the originators of the data be credited as authors in any new publication. The question arises, though, whether actual authorship is always appropriate. For example, the data collection team may not want to be held responsible for the secondary analysis, or may disagree with some of the points being made. An alternative is to require citing the

data set itself, which can be self-published with a digital object identifier (DOI). Such emerging practices are part of an increasing realization that traditional models of article authorship and citation are unfit for a science that, more and more, follows larger-scale modes of collaboration (e.g., Vasilevsky et al., 2021).

Bad-faith uses of open data are sometimes cited as a concern. Especially in controversial areas, some researchers fear that making data open would enable ideologically motivated critics to swoop on minor inconsistencies as a pretext to invalidate the whole research effort in the court of public opinion. This concern is real but can be addressed. First, bad-faith critics are by definition going to find other ways to tear down the research -- including pointing out that data are not available for inspection. Critics operating inside academia can and will use traditional means of data verification to harass researchers, and other tools such as freedom of information requests (Lewandowsky & Bishop, 2016). Selectively refusing their requests on the grounds of their assumed ideology looks bad. It also sets a precedent for researchers to refuse scientific requests that, while adversarial, are also legitimate.

Ultimately, though, the claim of science to be based in truth overcomes concerns about the privacy of the means used to arrive at that truth. If critics exploit minor mistakes in reports based on open data, then researchers in controversial fields have a special obligation to make sure those mistakes do not exist. This is not just to protect themselves from the activities of biased critics working outside the scientific paradigm, but to fulfill the special obligation of highly impactful research to have an unimpeachable basis in fact. Lewandowsky and Bishop (2016), writing from experience, give further valuable advice on dealing with scientific harassment while maintaining commitment to scientific and civic openness.

Finally, open data can require a substantial amount of work to meet standards of usability. The possibility of charging a fee under the traditional APA rules reflects concerns about this expense. If a laboratory's internal data set is rife with enigmatic variable names known only to their creator, with no record of how they were calculated, then preparing a fully usable dataset with codebook and analysis scripts will take considerable time and effort. However, there is another way to see this issue. Documentation is a corner that is often cut, but should not be, even if data are never shared. That is, a usable and documented data set helps its laboratory work efficiently and avoid errors. It also preserves the legacy of research if key personnel leave the project or suffer misfortune. Requiring open sharing of usable data and materials rewards those labs that are already following good internal practice at the cost of their time to do other things, while incentivizing less tightly run labs to catch up. Under this perspective, the APA policy of fees for data inspection should rightly be limited to the costs of reproduction and transmission, not the costs of catching up on what researchers should have been doing from the start.

## *New Rules?*

The presence of open data, materials, code, and explicit disclosure statements is easily checked when evaluating a manuscript before or after publication. However, further checking for completeness and reproducibility of data looks to be a task for a dedicated researcher or administrative person. Journals with a high rate of submissions may not be able to afford this level of care. Likewise, only when there is a special reason to look closely at the data can we expect readers of manuscripts to do this level of checking themselves. While open data and materials do not directly bear on the strength of evidence, they do indicate the researchers' willingness to allow

scrutiny of how they arrived at their conclusions, and to enable alternative analyses to be carried out.

As for disclosure statements, they add to a reader's confidence that the data being reported are a complete picture, rather than the most supportive subset of an unknown number of undisclosed measures and analyses. Because of the uncertainty around research without a disclosure statement, it is impossible to suggest an exact adjustment to p-values or other heuristics for evidence when considering it. Less formally, a reader might be more willing to accept conclusions from tests that are not consistently significant, or from p-values close to .05, if they were confident that selective reporting was not being done in the article.

## Evidence Solution 7: Enable Pre-Registration, Registered Reports, And Honest Exploration

### *Why?*

Early in the 2010's, as we have seen, unreported flexibility in data analyses was identified as a culprit in false-positives, failures to replicate, and underestimation of effect sizes. Researchers had been given freedom to wander in a metaphorical "garden of forking paths" (from the Borges short story of the same name; Gelman & Loken, 2014), trying out analyses until they stumbled on a positive conclusion. Complete disclosure solves some, but not all, of this problem of flexibility, because there are still many choices that can be made starting from a fully reported dataset and methods.

Instructive demonstrations of just how much the garden paths can fork have come from "Many Analysts" projects in different areas of psychology and other sciences. In these projects, multiple independent research teams receive the same data sets with instructions to answer a single question from them (e.g., Silberzahn et al., 2018 in social psychology and other social sciences; Hoogeveen et al., 2022, for religion and well-being; Botvinik-Nezer et al., 2020 for neuroimaging; Landy et al., 2020 in economic and social psychology). Teams have generally made widely divergent choices in analytic technique and other options (e.g., which covariates to include, which arguably bears on theory rather than method; Auspurg & Brüderl, 2021). These choices in turn lead to widely divergent conclusions. For example, in the first reported "Many Analysts" project (Silberzahn et al., 2018), the question was whether bias based on players' skin color influenced red card penalties within a single season of four European men's football (soccer) leagues. None of the 29 teams chose the same statistical analysis and set of covariates to answer the question. Conclusions correspondingly ranged widely, from no bias to large and substantial bias. Similar lessons can be drawn from the other projects, sometimes less dramatically.

The Many Analysts results visibly justify the concerns that might be raised about researchers selectively analyzing data to get their desired result. Put more formally, using p-values as inferential statistics requires precise specification of the research problem ahead of time (de Groot, 1956/2014). If ten procedures are tried and only the one yielding the lowest p-value is reported, multiple correction is necessary. But because the outcomes of these procedures are not independent and the degree of their correlation is not generally known, precise mathematical correction is not possible. Although one solution involves explicitly reporting a wide variety or "multiverse" of analytic methods (e.g., Del Guidice & Gangestad, 2021), preregistration has been more commonly used in

social psychology over the last decade. Preregistering analyses involves choosing and reporting the best procedure ahead of time, on statistical grounds that may rest on the properties of the data or of the question, but which are applied independently of the outcome.

Another issue with retrospective accounts of research was first observed by Kerr (1998) and dubbed HARKing, which stands for Hypothesizing After Results are Known. This phrase, though, is a slight misnomer. What is objectionable according to Kerr's analysis is not the mere fact of basing hypotheses upon exploratory inferences from data, which is a normal part of science. Rather, the practice he objected to was writing about exploratory results as if they had been foreseen. As with the selective use of analyses, HARKing makes the results look more impressive. But HARKing does so in a different way: not via the heuristic of strong evidence, but via the heuristic of a strong and simple narrative, presenting the appearance but not the substance of the requirement for results to be grounded in theory in basic-research psychology journals.

Online preregistration provides a way to show that a researcher has specified the relevant theories and their predictions ahead of time, as well as specifying their analyses. Preregistration makes available a document that is date-stamped to establish that it comes before data collection. While opinions vary about what should go into a preregistration (Ledgerwood, 2018; Nosek & Ebersole, 2018), at a minimum it should specify a plan for analyzing the data, including data preparations such as computation of measures and exclusion of data points. To ensure that any post-hoc additions to data collection are also recorded, a target number of participants and any procedure for increasing these (e.g., if some prove unusable) is also often recommended. These measures limit the possibilities for p-hacking and selective reporting of analyses. More maximal templates for preregistration identify the main and secondary hypotheses of the research, as well as other methodological features (e.g., Bosnjak et al., 2022); Giner-Sorolla & van t'Veer, 2016; Preregistration Challenge template as described in Toth et al., 2021). These features work to limit HARKing by making sure confirmatory analyses are clearly identified and separated from exploratory ones.

The Registered Reports article format is a stronger defense against publication bias, regulating not just selective reporting within a study but the decision to publish positive or negative results of the whole study (Chambers, 2013; Chambers & Tzavella, 2022). In this format, an article built around a preregistration is peer-reviewed before data collection begins. If the rationale and plan for the study are ultimately found acceptable, the editor commits to accept and publish the article regardless of the outcome of the study.

The Stage 1 submission in a Registered Report includes a conventional Introduction section, a full Methods section including planned participant numbers and recruitment method, and an Analysis Plan section that takes the place of the Results section. This plan undergoes peer review, including possible revisions, which can deliver the benefit of peer input on the methodology before committing to running the study. When the editor is satisfied with the manuscript or its revision, they make a decision of *in-principle acceptance*. The authors then run the study, after preregistering the accepted plan (Chambers & Mellor, 2018), and submit a Stage 2 report: a complete research article following the analysis plan and including a standard Discussion section based on the results. The acceptance decision commits the journal to publish the Report regardless of the outcome of the analysis. Grounds for rejecting the Stage 2 manuscript are few and rare: for example, if the authors deviated from their plan without justification, or if they insist on an unreasonable interpretation of their results in the Discussion section.

## *Uptake*

Preregistration in social psychology is currently common if not yet standard. On the Open Science Framework (osf.io), one major site for archiving preregistrations, we can find over 15,000 registrations of projects tagged "social psychology" alone. Although no major social psychology journal requires preregistration of non-Registered Report articles, several offer open science article badges for registration, and others encourage them as part of author submission guidelines. However, from 2015-2017, Hardwicke et al. (2022) found very little use of preregistration across psychology journals (3% of articles).

To examine more recent uptake of preregistration specifically in social psychology, we turn again to Google Scholar. Out of 1020 articles published by Journal of Personality and Social Psychology from 2016 to February 2024, 421 contained the word "preregistered" or "preregistration" in their text (excluding cited articles), a rate of 41.3%. Encouragingly, this rate is greater among articles published since the start of 2021, up to 68.5%. In a few selected other journals, the comparable rate is: PSPB, 37.7% (67.7% since 2021); SPPS, 39.1% (54.44% since 2021); European Journal of Social Psychology: 23.2% (43.9%); JESP, 44.0% (73.1% since 2021). Differences in these rates depend in part on editorial policy and the kind of research published, and not all studies in an article may be registered. However, preregistration in social psychology is clearly practiced by many. Its use is rising across the board, to the point of becoming a majority practice in many major journals.

The quality of published preregistrations, however, has come in for scrutiny. Researchers might cut corners with the new heuristic by submitting an underspecified preregistration, allowing a modest amount of p-hacking to happen. But what counts as underspecified? Templates and authors give varied advice on how much detail is needed; a particular gray area surrounds the reporting of contingency plans for violation of analysis assumptions, which can become very long indeed (Giner-Sorolla & van t'Veer, 2016).

Bakker et al. (2020) sampled an archive of preregistrations, finding that many of them still left considerable latitude for researcher decisions in analysis. Looseness in the preregistrations was reduced, if not eliminated, when using a detailed template rather than going free form. Further evidence that published links to preregistrations require quality checking comes from Claesen et al. (2021). Out of 38 early-adopting studies in *Psychological Science* whose papers earned a preregistration badge between 2015 and 2017, 24 had some form of undisclosed deviation from the registration, while 11 could not be evaluated in the first place because they were not clearly accessible, not clearly specified, or lacked a time-stamp. These results are confirmed by van den Akker et al (2023b), who focused on the match between hypotheses mentioned in preregistrations and in published articles from 2017 to 2019, and found that more than half either omitted or added hypotheses compared to the preregistration.

As for Registered Reports, at least ten social psychology journals offer the format, including one exclusively dedicated to it, *Comprehensive Results in Social Psychology* (Jonas & Cesario, 2016). In journals where Registered Reports are an option, uptake has been visible but slow. For example, at *JESP* in February 2024, out of about 800 articles published since Registered Reports began there in 2018, only 38 have been Registered Reports. A recent overview of journals offering Registered Reports across a number of disciplines finds a similarly low rate (Montoya et al., 2021).

Contrasting with preregistration, Registered Reports' potential to reduce publication bias is clear. Scheel et al (2021) compared Registered Reports in psychology to standard articles. The standard articles reported 99% hypothesis-supporting positive results, but the rate for Registered Reports was only 44%. This contrasts with more ambiguous results for preregistrations. A similar look at non-

Registered Report studies in psychology has found that preregistration alone shows little tendency to reduce the proportion of positive results or their effect size (van den Akker et al., 2023c).

Under the assumption that Registered Report studies had 80-90% power to detect a population effect of interest, the assumed proportion of true hypotheses in Scheel et al. (2021) has a 95% confidence interval between 32% and 68%, averaging just under 50%. As with the earlier discussion of the ideal rate of positive results in the literature, this rate appears healthy in a discipline that values research questions which are neither sure shots nor long shots. The only caution is that research ideas submitted as Registered Reports may not be representative. These studies might be more likely to select questions for which finding negative results would not just be meaningful, but also more likely, anticipating the guarantee of publication.

## *Costs And Limitations*

Writing preregistrations undeniably takes additional work, much of which is spent trying to work out the best analysis plan ahead of time. A recent survey of psychology researchers found that most viewed preregistration positively, but many also recognized that it increased stress and added work (Sarafoglou et al., 2022), echoing the feelings expressed by the president of the Association for Psychological Science in a much-commented editorial (Goldin-Meadow, 2016). Among another sample of researchers, Washburn et al. (2018) found that added work was brought up sometimes as a reason not to preregister studies. However, other reasons were far more frequent: the lack of journal requirements to preregister, or the exploratory or descriptive nature of some of their studies. The stress and effort of committing to a detailed plan might be one reason why some researchers submit poorly specified preregistrations, as we have seen. Consequently, additional work also falls on journal editors and reviewers to read the preregistration and make sure the report matches the plan.

Sarafoglou et al. (2022) also observe that some advocates of preregistration claim it is easy, while others acknowledge its difficulty. This split reflects contrasting approaches to preregistration. The first approach takes as its motto, "preregistration is a plan, not a prison" (Nosek et al., 2019). In this view, preregistration works mainly to increase transparency of the research process, and researchers may deviate freely from their plan, if they signal and document what they are doing and why. The other approach takes the implied opposite view: preregistration, if not hyperbolically a "prison", is at least a strict way to regulate analysis in order to meet the formal requirements of hypothesis testing. Deviating from a preregistered plan undermines that purpose, unless part of the plan can be shown outright to be wrong.

The "prison" approach inspires frequent objections in social sciences, on the grounds that preregistration inhibits exploration and creativity, or is unsuited for descriptive and exploratory research (McDermott, 2022; Pham & Oh, 2021). The "plan" approach allows more freedom for these things and underlies several proposals to apply preregistration outside the traditional context of hypothesis-based inference and experimental trials, even to qualitative research (Haven & De Grootel, 2019). The only question remaining, though, is why the "just a plan" kind of preregistration is necessary at all. More specifically, if the point of preregistration is simply to document the research process, there is no need for it to be done ahead of time ("historical transparency") instead of while the research is going on ("contemporary transparency"; Rubin, 2021).

Devezer et al. (2021) further argue against the restrictive application of preregistration and make a case for formal quantification of the risks being taken when adjusting the sample or analysis after

the fact. In their view, the insights that can arise on contact with the data, when properly corrected for, outweigh the need to specify everything ahead of time. The practice of *sequential sampling* takes a more specific but similar approach (e.g., Lakens & Evers, 2016). Pre-registration usually fixes a sample size ahead of time, which is often based on uncertain estimates of the population effect size via power analysis. Sequential sampling instead seeks information about the effect size as the research proceeds. It allows multiple smaller samples to be added if initial inferential tests are inconclusive, assuming that p-values are appropriately adjusted for these multiple non-independent tests of significance. While adding more participants after seeing non-significant results is a form of p-hacking if not disclosed (Simmons et al., 2011), it is a legitimate practice if disclosed and corrected for.

Another class of critique points out that preregistration is not necessary or even useful for research based on a well-specified theory (Fiedler, 2018; Szollosi & Donkin, 2019; van Rooij, 2019). Under strong enough theories there is no need to reiterate what the focal hypothesis tests are or what they mean (that is, HARKing is not possible). Further, if the theories follow on through well-developed methods, then the correct technique of analysis has been established upon beforehand and needs no prior specification. However, as we will see closer to the end of this chapter, neither theory nor methods in social psychology currently stand at this level of development.

Preregistration, in sum, looks to be an adequate tool to control the undisclosed flexibility that has led to several failure cases in the standard mode of social psychological research. Although it is less convincing as a universal necessity for good research, social psychology is currently not well equipped either to precisely quantify risks from analytic flexibility, or to implement the kind of theory that marks out but a single analytic path in any given paradigm. Further, preregistration may be partly supplanted by other practices considered in this chapter, such as disclosure. Registered Reports, on the other hand, additionally tackle publication bias at the root by separating evaluation of the ideas and methods from knowledge of results. Perhaps this is why Registered Reports have not come in for as much criticism as preregistration has. In fact, for those who prefer the benefits of Registered Reports without the restrictions of preregistration, *results-blind analysis* comes to the rescue (Dutilh et al., 2019; MacCoun & Perlmutter, 2015). This technique lets researchers avoid bias towards seeking positive results by concealing the conclusions of data analysis until all decisions have been made; for example, by masking the labels used, or working out optimal methods on a smaller subset of the data that may not give the same conclusion.

Registered Reports have many benefits, including being able to modify the research in response to peer review comments, and having a guarantee of publication regardless of results. Yet their limited uptake hints at perceived downsides outweighing all these positive points. Chambers et al. (2022) notes that the special timeline of Registered Reports can form a barrier to working them into a laboratory's research cycle. Specifically, the wait between finalizing the study and running it depends on the amount of time the peer review process takes, involving revisions and indefinitely long waits for editors and reviewers. A wait ranging at random from two months to a year is not very attractive to graduate students and postdoctoral researchers on strict timelines. Nor does it fit well with time-restricted opportunities to collect data on specific events or populations. At a minimum, editors should streamline Registered Reports by evaluating any revision themselves rather than sending it out for reviewers' second opinions. More profound measures have been implemented at some journals, such as *Royal Society Open Science* and *Nature Human Behaviour*, who promise an expedited review to Registered Reports rendering a decision in no more than four months (Chambers & Tzavella, 2022) -- surely a good incentive to submit using the format, even leaving aside the benefits of Registered Reports themselves.

## *New Rules?*

After this overview, it should be clear that the mere existence of a preregistration is no guarantee of strict adherence to an a priori plan. Studies that claim preregistration through their title, abstract or badge often betray the principle of preregistration by straying in undisclosed ways from the plan. A minimal quality standard would require that preregistrations at least report all results according to the original plan. However, vague plans and disclosed deviations (such as from unexpected circumstances in the research or errors in the registration) are harder to factor into the evaluation of preregistrations. In an encouraging sign, some authors have recently developed guidelines for best practices in reporting deviations from preregistration (Willroth & Atherton, 2024), distinguishing between necessary improvements and results-biased changes that dilute the registration's rigor.

It has been argued that even vague preregistrations are better than none because they constrain analytic freedoms in some way (Simmons et al., 2021). But there are also reasons to treat preregistration as a guarantee of a certain level of commitment to confirmatory analysis -- a marker of confidence that a study's p-values were not produced after an unprincipled process of discarding exploratory analyses. Allowing preregistrations to be vague dilutes their value as a marker. So does insisting that preregistrations be extended to exploratory and descriptive research, not to mention qualitative research, which has a long tradition of using its own techniques to ensure transparency and address bias from the researcher's standpoint (e.g., Madill & Gough, 2008).

As for deviations, if new hypotheses are generated after the fact, they should be treated as exploratory and tentative in the Discussion and Abstract and should not be the basis of the study's headline finding as announced in the title. Deviations involving the analysis plan are another matter. They are sometimes inevitable, due to realizations after the fact or emerging properties of data, as Willroth and Atherton (2024) recognize.

But disclosed deviations present a special danger of interpretation. Statisticians often disagree with each other, to put it mildly, and psychology researchers vary greatly in their adoption of techniques from statistics. So, a wide range of hypothesis tests, assumption checks, and corrections can be justified after the fact as statistically correct in the eyes of someone, somewhere. Was the post-hoc change carried out because it gave better results from the authors' point of view, or because it was the right thing to do statistically? The ambiguity surrounding this question can be resolved, perhaps, by reporting both originally planned and deviating analyses, as Rubin (2021) suggests. If by coincidence the analysis chosen looks better for the researchers' stated hypothesis, then the burden is on the authors to convince readers that statistical choices were made on principled grounds, and that alternative analyses are not justified. Admittedly, it is hard to make such an argument after the fact. This should lead authors to spell out contingency plans for the most usually violated assumptions in their pre-registration.

Registered Reports grant a stronger guarantee of the principle that good studies building on good ideas should be published regardless of their results. Preregistration has been seen as an attempted remedy for social psychology's deficiencies in theory and method specification, by forcing these blanks to be filled in ahead of time. Registered Reports go further and evaluate articles on strong theory and methods from the start. With a hypothesis based on a tightly specified theory, together with validated, high-powered methods that minimize the risk of false-negative results, reviewers can be sure that a negative result counts as evidence against the hypothesis, as much as a positive

result counts in its favor. Under these conditions, there are few problems with a negative result being allowed into the literature.

## Evidence Solution 8: Strengthen Theory

### *Why?*

Although much attention in the early 2010's was devoted to statistical, methodological, and reporting solutions to false-positive failure cases, many of these failures also took place in a vacuum of theory. Bem (2011) made no claim to explain the precognitive phenomena he claimed to have shown, which would have required a thorough revision of fundamental axioms in neuroscience and physics. Another instructive case focuses on experiments in the "behavioral priming" paradigm (or "social priming"; but that term may not be completely accurate, see Sherman & Rivers, 2020). Following the lead of Bargh et al. (1996), this line of research has sought to demonstrate that subtle but consciously perceptible cues in the environment have appreciable effects on behavior, judgment, and motivation. Whether these cues activated cognitive categories, motivational mindsets, or embodied sensations, the behavioral priming paradigm was very productive and visible during the first decade of the 2000's (Strack & Schwarz, 2016).

However, behavioral priming was also implicated in the replication controversies that followed, starting with an early failure by Doyen et al. (2012) to replicate the best-known finding from Bargh et al. (1996, Study 1): priming the category of elderly people led young adult participants to walk more slowly. In the discussions that followed, it was not at all clear what conditions could either guarantee or falsify the workings of behavioral priming. The phenomenon seemed to be susceptible to context effects ranging from the fundamental (participants in Belgium having different associations of words to the elderly than participants in Manhattan) to the arcane (differences in the layout of the walking corridor influencing participants' attention). The related field of embodiment research, too, was described even by its proponents as "more descriptive than explanatory" at the end of a decade of research (Meier et al., 2012, p. 705), encouraging further examination of boundaries and preconditions.

Some experiments in the behavioral priming tradition indeed have shown an atheoretical, cavalier approach to specifying the basis for their implicit influences. For example, Leung et al. (2012) published an article in *Psychological Science*, then the highest-impact outlet for original empirical social psychology research reports. In this article, across five studies, participants who physically acted out the metaphors "think outside the box," "on one hand, then on the other hand," and "put two and two together" showed greater creativity (at least on some measures and trials) than control participants did. However, the claim that all these metaphors related to creativity rested on a single sentence: "Such prescriptive advice is common to people in research labs, advertising teams, the halls of higher education, and other contexts in which pioneering, novel approaches to pressing problems are valued." (p. 502). While the box metaphor is well-established in usage, it is not immediately apparent how being evenhanded, or solving a simple and obvious math problem, might relate to more creativity, instead of limiting creativity or being unrelated to it. Evidently, the editor and reviewers were more persuaded by the positive results of these studies, than by any groundwork establishing why the experiments were a good test of the embodiment theories cited. The studies, like numerous other articles of the period, lacked theoretical reasoning about why specific metaphors should be expected to evoke specific outcomes.

Cesario (2014) warned that priming effects should not be expected to work universally without further specification in theory. Non-replications as well as successful conceptual replications would be inconclusive without a falsifiable way of telling when priming and related effects are expected to work and when they are not. For example, the accessibility of the concept, the strength of the link between two concepts, the exact social nature of the relationship being activated, and its interaction with chronically and acutely accessible motivations -- all these need to be established before making a firm prediction about a specific priming effect. Although some of these more rigorous theories of priming were in print by 2010 (see Cesario, 2014, note 3, p. 46), not all published priming research has drawn on them. Apparently, at one point it was enough to meet the sophistication heuristic in publication by referring to an associative theory of activation, and to meet the novelty heuristic by demonstrating a surprisingly strong effect of a non-obvious association on a behavioral or expressive outcome.

Another topic where calls for better theory have surfaced is ego depletion, a phenomenon in which executive mental functions, after being used in an effortful task, are weakened in a subsequent task (Baumeister et al., 1994, 2000). The central metaphor of ego depletion involves a limited reserve of self-control that is drawn upon and subsequently reduced. However, since the idea was initially proposed, despite hundreds of experiments reported in support, there have been several large-scale replications and bias-corrected meta-analyses that give less impressive accounts of the existence and strength of the phenomenon (see Inzlicht & Friese, 2019 for a summary).

Today, there is still little consensus about the replicability, conditions, and processes of ego depletion effects. Lurquin and Miyake (2017), in addition to critiquing the loose methodology in depletion research, also call for a better specification of the central theory -- in particular, quantifying how and when a given task will sap the proposed reserve of ego-strength. Indeed, some research programs have dedicated themselves to specific explanations of ego depletion, such as the now-discredited blood glucose idea (Gaillot et al, 2007; Kurzban, 2010) or the possibility that depletion works through self-perception or motivation rather than loss of cognitive capacity (e.g., Inzlicht & Schmeichel, 2012; Job et al., 2010). But the calls for more theoretical development emphasize that these explanations need to be central to the theory itself.

Some writers have called for more precise and better-defined theory development across the board in social psychology, or in psychology more generally. Klein (2014) made an early point reiterated by many of these critics: to test their truth, theories must be designed with enough specificity to identify conditions that will falsify them if results are convincingly negative. To quote Scheel's (2022) titular argument, most research findings in psychology are "not even wrong" because their claims are not specified in enough detail to extend to other tests of a common theory. Fiedler (2017) likewise argues that strong psychological theories should draw from well-established principles such as psychophysical or economic laws. He adds, however, that these investigations should leave room for less conclusive forms of question-asking in psychology, such as model fitting.

Lewandowsky and Oberauer (2018) agree with these points in general but make a specific case for the utility of models. They demonstrate through mathematical simulations how ideas built on strong theory do not require many of the corrective measures proposed for the kind of weakly theoretical experiments more characteristic of social psychology, such as preregistration (see also Fiedler, 2018). Van Rooij and Baggio (2021) focus more specifically on what our theories are about. They maintain that a literature that consists entirely of artificial experimental settings will offer little clarity as to why effects come about or what other effects we can expect to see. Instead, they advocate a focus on building plausible explanations of natural human capacities using a

systematized, computational approach to theory building taken from Marr (1982/2010). To cite just one more of these calls to theory, Muthukrishna and Henrich (2019) emphasize the importance of having theories that are not just well-developed, but expansive enough to be able to unify efforts in a field. They give as an example dual-inheritance evolutionary theory.

But the theory problem can lead in a different direction: if social psychology is not yet developed enough to generate adequate theories, perhaps we should grant greater recognition to descriptive and exploratory research. Rozin (2001) made an early entry in this category, advocating that social psychologists should accept and carry out the kind of descriptive and exploratory research that characterizes other sciences without a strong theoretical basis. He cited Solomon Asch's demonstrations of conformity and group processes as an example. Scheel et al. (2021) more recently have reiterated this call for the fuller development of the concepts, causal relations, and boundary conditions (as well as methods) underlying psychological theories before they are ready to be put to a confirmatory test. These goals can be achieved through conceptual work supported by validation research, as well as the kind of description and exploration advocated by Rozin.

Whether the call is for theory to get tighter, or for research to get looser, one common point in this criticism is that conceptual priors for our evidence should be better established, even if they cannot be easily quantified. In formal models of null hypothesis testing, we have seen that the prior probability of the proposition counts for more in establishing the credibility of the result than does the statistical power of the test. This does not mean that credible research ought to test dull and obvious propositions. Rather, theory can validate a superficially surprising finding by providing a careful chain of argument showing why it should *not* be surprising -- the difference between a magic trick that is explained, and a magic trick that is not.

By the prevailing rules of publishing, it may seem unfair to evaluate evidence more harshly if it is not supported by careful and plausible theorizing. However, Bayesian logic supports the motto that "extraordinary claims require extraordinary evidence." Laypeople, too, understand that more plausible findings are more likely to be replicated (Hoogeveen et al., 2020). The challenge for social psychology is to separate well-developed theories from ideas that only specify effects rather than explanations. There is room for both modes of research in the practice and reporting of our science. The mistake has been to require authors to pass off the looser research mode as the tighter one as a necessary sign of sophistication, encouraging the HARKing behavior identified by Kerr (1998).

## *Uptake*

There are obvious challenges to determining whether theoretical framings in social psychology have improved over the past ten years. Theory's contribution to a research article must be judged qualitatively and subjectively. There are no widely accepted bibliometric marks of stronger theorizing as there are for preregistration, replication, or open data. Theory considerations have not usually appeared in surveys of research practices among psychologists (e.g., Banks et al., 2016; Motyl et al., 2017). One study of articles in *Psychological Science* from 2009-2019 used mentions of theory as a proxy for theory development and, by this standard, found that most articles were not presented as tests of theory and almost half did not even mention it (McPhetres et al., 2021). Also, most mentions of theory were not repeated across different articles, supporting Mischel's observation that theories are seldom shared across research labs in psychology.

Some journals, nonetheless, have increased the importance of theory in their submission guidelines since 2010, perhaps inspired by discussions of theory as a positive factor in the credibility of

evidence (e.g., Rothermund & Koole, 2020). As a detailed example central to social psychology, the Attitudes and Social Cognition section of *JPSP* since 2017 has required Discussion sections to include, and highlight in green within the manuscript, some words about theoretical implications. This requirement was inspired by concerns about a field that has sometimes privileged atheoretical "sexy" results to the detriment of robustness (Kitayama, 2017). However, the *JPSP* theory requirement is framed in much more general terms than the ones advocated by recent authors. As a result, it is easy to see this requirement being satisfied simply by using one or more of the pseudo-theoretical surrogates identified by Gigerenzer (1998; e.g., mere labeling of single or dichotomous constructs) rather than by a formally specified theory. The policy might curb blatantly atheoretical research, but it does not address the larger concerns that have coalesced around social psychological theorizing.

## *Costs And Limitations*

In the past, theory's role in psychology has sometimes been viewed askance. For example, Greenwald et al. (1986) posed the provocative title question, "Under what conditions does theory obstruct research progress?" The answer assumed that researchers, in seizing on a theory and proceeding with research designed to confirm it, would ignore disconfirming results. In their words, the problem occurs "[w]hen the researcher repeatedly resolves the disconfirmation dilemma by retesting the prediction rather than reporting results" (p. 220). Presumably this objection is not the fault of the theory itself, though, but of selective reporting. Indeed, the authors follow on by calling for more complete reporting, scientific self-correction in the literature, falsification testing, and theoretical approaches that directly address boundary conditions -- remedies to the problem of evidence which now are very familiar.

Another drawback to stronger theory might be the restriction of research to testing *a priori* well-specified questions. The drawback can be solved, however, by also valuing work that raises interesting questions in search of theoretical explanations. Credit and recognition could go, for example, to three kinds of research team. First, the ones who came up with the idea and reported initial descriptive and exploratory studies; then, another team who did important work filling in the theory and testing the boundaries of the effect; and ultimately, a third team who took the research question to a strongly theorized culmination. This scheme would allow all kinds of contributions to count -- creative thought and keen observation at first, perceptive methodology development in the middle, and strong quantification at the end. Of course, the equality of these three modes would have to be supported by active and explicitly advertised editorial policies, countering existing prejudice against descriptive work and recognizing the difficulty of constructing strong theory.

## *New Rules?*

Unlike calls for improved reporting transparency, calls to improve the use of theory have not been accompanied by a single, easily verifiable change in practice. Directed acyclical graphs to explicitly model causal possibilities and alternatives (Grosz et al., 2020; Rohrer, 2018), formal processes of elaboration and testing (Borsboom et al., 2021), and systematic visual methods (Gray, 2017), have all been proposed. Formal computational modeling also has its advocates, as a way to improve and standardize the specification of theories (Grahek et al, 2021; Guest & Martin, 2021; Scheel, 2021). The suggestion to model theoretical assertions computationally in fact predates the current reform movement by some years (Hastie & Stasser, 2000; Smith & Conrey, 2007).

But by and large, it is not clear that improvements in theory need to follow any one signpost. Favoring articles that present strong theoretical tests depends on recognizing when a strong theory occurs, as opposed to just an idea. Evaluators also need to relax the heuristic that good research should appear to be novel and surprising. Shallowness, disjointedness, and redundancy in theory occur in part because researchers get the most respect for claims they present as entirely novel. But in mature research disciplines, truly new theories are rare; when introduced, they are major events inspiring research by dozens of labs. "Incremental" should be a desirable feature of theoretically grounded research, not a term that editors use to excuse their desk rejection of a manuscript.

# CHALLENGES TO THE RULES OF MANIPULATION AND MEASUREMENT

Criticism of the ways social psychologists present evidence has been echoed in recent years by a less audible drumbeat of criticism about their methodology. Concerns over methods are aptly discussed immediately after concerns over theory. As we have seen, the wish-list for better theorizing in psychology has included tighter definitions of constructs, which would lead to clearer operationalizations of manipulation and measurement (see Grahek et al., 2021). Indeed, to quote the title of a commentary by Greenwald (2012), "there is nothing so theoretical as a good method." That is, when connecting effects to explanations, it's crucial to specify which part of the method is essential to testing a theory and which part is incidental. There is also a growing realization that meta-analysis cannot give good estimates of effect size if it covers greatly discrepant methodologies whose weight in the analysis varies arbitrarily with their prevalence in the research literature (Elson, 2019; Linden & Hönekopp, 2021). The average effect sizes that result from such a meta-analysis may be no more valid than the average weight of all the animals in the zoo. Standardizing well-validated methodologies, then, and accounting for method interactions, can lead to more accurate estimates of the size as well as direction of effects.

## Why?

In recent years, there have been noteworthy critiques of the low priority placed on accurate and valid construct measurement in various subfields of psychology. One of the more prominent critiques has targeted "questionable measurement practices" in psychology, calling out the "measurement, schmeasurement" attitude in the field (Flake & Fried, 2020). To concerns about undisclosed flexibility in measurement decisions, such as picking which items from a scale to use, they add further concerns about poor reliability and unvalidated scales. From studies of measurement practices in psychology (e.g., Hussey & Hughes, 2020), Flake and Fried conclude that psychometric concerns are routinely ignored in research and publishing. Weidman et al. (2017) issued an earlier and more specifically targeted brief against imprecise and ad hoc scaled measurement in emotion research. Schimmack (2021) has warned of a deficit in construct validity in psychology, again uniting theory and method concerns, which can best be solved by carefully mapping constructs using nomological networks.

Together with the call for improved measurement has come a focus on improving rigor in experimental manipulation. Over the past few decades, positive results from an experiment have usually been taken as sufficient evidence that a manipulation worked. But things become more complex when independent teams try to replicate those results, or when a research team submits a

Registered Report that might be published with null results. How can we be sure that a negative result reflects a meaningful near-zero effect, rather than a simple failure to set up the experiment correctly? One answer is to include manipulation checks: that is, measures of the independent variable which should vary appropriately between experimental and control groups. This construct-validating definition of the term "manipulation check" should not be confused, as it commonly is, with more simple checks of attention or understanding (Ejelöv & Luke, 2020).

The usefulness of manipulation checks in experimental psychology has been under debate for some time now. Some authors vouch for their usefulness and necessity. For example, Lench et al. (2014) stressed the importance of establishing an effective manipulation in social cognition research, and recommended analyses to test whether the manipulation check in turn predicts the outcome. Citing previous writings on validity such as Campbell and D. T. Fiske (1959), Fiedler et al. (2021) have argued that manipulation checks are necessary to resolve doubts about significance testing in the present era. They establish validity in both a convergent sense (testing for the presence of the variable being manipulated) and a divergent sense (testing other variables to rule out their influence).

However, other authors have had their doubts. Sigall and Mills (1986) and later Fayant et al. (2017) have argued that measures within the same session cannot by themselves establish or rule out the validity of an experimental manipulation. Kidd (1976), and later Hauser al. (2018) as well as Fayant et al. (2017), also caution that self-report checks can themselves influence the procedure by making participants aware of the variable being measured. They suggest nonverbal checks as better alternatives. Gruiters (2022) contends that manipulation checks per se are unnecessary tests that only inform us about effects at an operational level, and can undermine inferences about the underlying constructs. Similar to Fiedler et al., (2021), Gruiters recommends a broader focus on a manipulation's specificity by assessing which of several different candidate constructs it influences most strongly.

Ejelöv and Luke (2020) review the controversies and end up cautiously advocating manipulation checks. They can be useful both as a test of experimental validity and as part of the measurement of causal strength. But attention must be paid to their effects on the experimental procedure, as well as to checks' own reliability and validity. Indeed, even when expressing doubt about manipulation checks, most of these authors also agree that manipulations' validity should be tested in some way. They suggest, for example, pilot tests conducted in separate sessions or with a different sample.

## Uptake

When looked at carefully, it seems indeed that articles in social and personality psychology need only address reliability and validity concerns in limited and rote ways to be published. Flake et al. (2017) reviewed articles in *JPSP* from 2014 and found that, while Cronbach's alpha was routinely reported, evidence about other dimensions of validity was absent from most measures. Hussey and Hughes (2020) examined a large data set with 26 commonly used scaled measures in social and personality psychology. While most of them did show good internal reliability through item-total correlation-based indices, which were commonly reported, other measures of reliability and validity fared less well, and only one of the 26 scales could claim that all psychometric properties were at the acceptable level. The prevalence of the Cronbach's alpha reliability metric in social and personality psychology, too, is at odds with statistical developments that offer better alternatives (e.g., Crutzen & Peters, 2017). Cortina et al. (2020) identify many of the same problems with

measurement in the Journal of Applied Psychology, offering examples and a road map for better scale development.

Looking at manipulations, Hauser et al. (2018) showed that manipulation checks were reasonably common in a selection of social psychology journals from 2015-16. Prevalence rates were in the 40-60% range, but actual prevalence among experiments was bound to be higher, keeping in mind that these authors did not apparently code only experimental studies. However, manipulation checks often showed problematic traits, such as lack of the counterbalancing needed to test the checks' influences upon and from the other variables in the study. Chester and Lasko (2021) focused more precisely on experiments published in *JPSP* in 2017 and found that about 40% of experimental manipulations were produced ad-hoc without any check or pilot testing for validity, with similar problems to those Hauser et al. (2018) found in the execution of manipulation checks. Ejelöv and Luke (2020) further found that many measures presented as "manipulation checks" in social psychology journals could not fulfill that methodological function because they were mere attention or comprehension checks that only determined whether participants could re-identify the surface meaning of stimuli.

The methodology critique seems to have gotten off the ground later than the false-positive critique. It has not seen the same kind of revolution as the rules of evidence have experienced -- even though, as with challenges to evidence standards, many of the measurement critiques existed many decades ago. It is true that some editorials have mentioned measurement problems . But as of today we are not generally seeing concrete policies being put in place at social psychology journals to require more careful development of measures and manipulations. Indeed, few papers in social-personality psychology mention problems with construct validity in their Limitations section (B. Clarke et al., 2023). Contrast this with the way open data, disclosure statements, power analysis, and effect size reporting have become *de rigeur* over the past ten years. What can account for the lag in methodology concerns?

Currently, critiques rest on claims that social psychologists' methods are inaccurate, unreliable, and noisy; in other words, that current practices are too likely to produce false negatives. The problem with such arguments is that they do not directly lead people to doubt positive results that have already been accepted as true. In fact, they might even lead to inaccurately seeing positive results obtained under bad conditions as "superpowered" in their ability to punch through the noise (Loken & Gelman, 2017). Then, add the belief that false-negative findings are primarily a cost to be borne by the research laboratory itself. With that understanding, the tradeoff between more careful methods and an increased rate of usable positive findings becomes a personal decision, but not a reason to doubt the quality of positive findings as presented.

Such a laissez-faire approach, though, can be refuted by direct analogy to advances in sample sizes. In fact, increased reliability, like increased sample size, feeds into statistical power. Larger effect size estimates are possible when reliability of variables is high, and effect sizes are an important input into assessing statistical power. Heo et al (2015) show how increased reliability increases statistical power. For example, in a paired t-test, using various combinations of parameters, increasing the reliability of the dependent measures from .70 to .90 increases the power of the significance test from the vicinity of .63 to the vicinity of .99 -- the equivalent, if testing a medium-sized population effect, of going from 23 participants to 81 (Heo et al., 2015, Table 2). All the arguments that low-powered research leads to a higher rate of positive findings that are false apply also to low-reliability measures. The measurement revolution may yet spark reform by finding a critical failure case among the inaccurate methods of studies that have proved hard to replicate.

There are also compelling cases supporting the other arm of the methods critiques -- not just that poorly tested methods are noisy, but that they might be measuring the wrong thing. For example, in research on ego depletion, finding replicable results and agreeing on what they mean for the theory has been made difficult by uncertainty about how to manipulate and measure key constructs (Forestier et al., 2022; Friese et al., 2019). To give one instance, ego depletion research has sometimes operationalized self-control by seeing how long participants persist in solving an unsolvable task (e.g., Baumeister et al., 1998; Dvorak & Simons, 2009). Yet it is not immediately clear which choice requires more executive control: mindlessly persisting at a task despite increasing evidence that the solution can't be found? Or keeping on despite mounting fatigue? As critical as these questions are, they are currently confined to influencing single areas of research, rather than being taken as a more general caution for social psychology.

## Costs And Limitations

There is a clear, prevailing reason why social psychologists do not generally put more effort into testing and validating methods. That process is rightly seen as effortful, time-consuming, yet little rewarded. Article word limits, and the limited patience of editors and reviewers, often preclude a full description of development and validation for newly minted measures. As opposed to personality psychology journals, which regularly publish scale and other method validation articles, social psychology journals do not often explicitly encourage stand-alone writing about methods development in specific research topics. Indeed, such articles might fall afoul of journal requirements to advance theory. As bibliometric research shows, the sheer number of ad hoc measures and manipulations is not visibly matched by an explosion of separate validation studies in print.

Journals that started rejecting any research for using ad hoc, poorly validated methods, would face a kind of trouble which the past decade's changes in reporting requirements did not have to face. The time scale and process of research would have to change to meet tighter methods requirements, leaving the most recently produced research with an uncertain publication future. And, if just a few journals at first tightened their requirements, they might find authors deserting them in favor of other, friendlier shores.

Methodologically careful research is undoubtedly better research. It certainly pays its laboratory dividends with a higher rate of positive and definitive findings. But with career pressures in academia being what they are, developing researchers cannot risk taking the slow path if this means they will end up with a superficially less impressive record than the competition. In this situation, adopting tighter methodology is a necessary accompaniment to other reform measures such as replication, Registered Reports, and publishing null findings. But while a coordinated advance in methodology would improve the field, it is still an unappealing prospect for most individual researchers. Changing social psychology's norms of measurement and manipulation would need strong targeted incentives.

## New Rules?

Flake and Fried (2020, p. 459) offer a series of concrete heuristic questions for evaluating the quality of measurement in any given article. They ask an author to: 1. define their key constructs; 2. justify measure selection and report previous validation; 3. match the measure to the construct; 4. describe

the measure's response coding, calculation and psychometric properties in the present sample; 5. describe and justify any modifications; 6. justify and report validation (or lack thereof) for new measures created for the research. Together with a clear preference for research using validated and justified measures, this checklist could be advertised as a set of submission guidelines in a journal. With minor modification it could also be applied to manipulations. A "validated measurement" or "validated manipulation" badge could be awarded as the first stage in recognizing this kind of effort.

# CHALLENGES TO GENERALIZABILITY

## Why?

Doubts about the focus of most social psychological research upon a narrow range of samples and settings are not new. Early critics noted the increasing reliance on university student samples in the 1950's and 60's, with concerns about what conclusions could be drawn from such a limited range of human culture and experience (Borgatta & Bohrnstedt, 1974; Sears, 1986). Others questioned why basic laboratory-based research seems to be glorified at the expense of real-world relevance (e.g., Silverman, 1971). Still others have decried reliance on easy self-report attitudinal measures at the expense of more effortful but realistic observations of behavior (Baumeister et al., 2007; McNemar, 1946).

These critiques have resurfaced recently as part of a current that often flows independently of attempts to reform evidence standards, and sometimes produces difficulties for reform solutions. The acronym WEIRD has been coined, for Western, Educated, Industrialized, Rich Democracies (Henrich et al., 2010). This label has helped to focus attention on the cultural limitations of most psychology research participants, including social psychology (but see critiques of the acronym's scope from Ghai, 2021, and Syed and Kathawalla, 2021). More general biases in the cultural and ethnic background of psychology as a discipline have been criticized, extending to the kind of research questions asked, the kind of answers received, and the people involved in research from participants to researchers to journal editors (Adams et al., 2015; Smith & Bond, 2022).

These critiques question the implicit distortions of a system that places the USA (and sometimes Canada) at the center, with majority-White English-speaking countries just outside (UK, Australia, New Zealand), followed by non-Anglophone Western Europe, then Eastern European, East Asian, and Global South countries. Looking further within countries, rural as opposed to urban people are understudied (Smith & Bond, 2022), not to mention people living outside of agrarian or industrial society entirely (Henrich et al., 2010). Within more central countries in this scheme, too, ethnic and racial representation among researchers, participants, and research topics has been called into question (Roberts et al., 2020).

Recent meta-scientific work identifies some mechanisms and consequences of these biases in the social sciences. Researchers, journal editors, and research populations are overwhelmingly drawn from more central nations and cultures in the above scheme (Arnett, 2008; Rad et al., 2018; Roberts et al., 2020; Thalmayer et al., 2021). This inequality supports an implicit bias in which research on people from central populations is seen as normal and automatically generalizable, while more peripheral groups are seen as special cases in need of labeling and explanation (cf. Hegarty & Bruckmüller, 2013; Rad et al., 2018; Syed, 2020). Kahalon et al. (2022), Cheon et al. (2020), and Castro-Torres and Alburez-Gutierrez (2022) each report bibliometric research showing that studies using

samples from more central populations are less likely to signal their origin in an article's title. Kahalon et al. (2022) further find that this discrepancy leads to perceptions of peripherally-oriented research as less important. Moreover, one recent survey of social psychologists from around the world unearthed accounts of pervasive discrimination against those working in more peripheral zones: being told their research is too specific, uninteresting, or needing proofs of validity that central-zone research does not need (e.g., control groups from a more central population, special validation of measures; Bou Zineddine et al., 2022).

Although few articles make explicit claims of generality to all humankind, psychology often makes these claims implicitly. Some subfield exceptions noted by Henrich et al. (2010) such as the study of personal values, emotional expression, and personality traits, all inherently focus on individual difference and thus allow for cultural variability was well. But overall, as Henrich et al. (2010) put it, "A typical article does not claim to be discussing 'humans' but will rather simply describe a decision bias, psychological process, set of correlations, and so on, without addressing issues of generalizability, although findings are often linked to 'people'" (p. 62). These observations receive quantitative support from Rad et al.'s (2018) analysis of articles in the journal *Psychological Science*.

Even without any talk of "humans" or "people", the pragmatics of language when reading claim, "X predicts (or causes) Y," lead to a generalized interpretation under Gricean norms of communication, unless further specified (Cheon et al., 2020; for further evidence drawing on cognitive linguistics, see De Jesus et al., 2019). Western readers who are still doubtful should ask themselves how they would react to research-based statements such as "The color red is implicitly associated with weddings" or "Showing the sole of the foot to another person provokes anger." If the reader thinks, "Not in my experience," perhaps wishing that the findings were more explicitly placed in context (of East Asian wedding traditions, and the manners of many Asian and African countries, respectively), then they are experiencing the other side of inappropriately generic statements of effects.

Recent critiques of the narrow focus of psychology are more far-reaching than before. Some argue that psychology has been too quick to label its results as general. They call for more cautious advances until generality is established, not just across human populations, but across classes of settings and stimuli too (Yarkoni, 2022). Other critiques move beyond concern about the generalizability of research, calling more radically for a reversal of flow, in which mainstream psychology takes seriously the ideas and priorities of more peripheral fields and regions. The decolonial psychology movement underscores that non-Western nations and cultures are not merely objects of study, but deserve to participate in science and contribute psychological ideas indigenous to their cultures (Adair, 2006; Decolonial Psychology Editorial Collective, 2021; Reddy & Amer, 2023).

Longstanding concerns about the relevance of social psychology research have also resurfaced in a more pointed form, in particular concerns about the privilege granted to basic research for implicitly treating its settings and study populations as context-free. Recent critiques have not just asked for more consideration of applied settings in publishing and funding. They have also asked to balance the process championed by Kurt Lewin (1951) at a critical stage of social psychology's development, of crucially incorporating field and applied findings in the development and assessment of basic theory, as well as exporting theory into practice (Berkman & Wilson, 2021; Giner-Sorolla, 2019). Also recently, there have been renewed cautions to hold back on applying basic research findings to policy problems such as those created by the COVID-19 pandemic, until

more is known about their generalizability across cultures, settings, and populations (Bryan et al., 2021; Cesario, 2022; IJzerman et al., 2020).

Any author who has been rejected but asked to resubmit the manuscript to a "more specialized" journal has witnessed the use of generality as an element of prestige in the publishing process. Indeed, the most prestigious journals are those which cover multiple subfields of psychology, or fields of science; those which require theoretical advances that increase the apparent generalizability (and generativity) of findings; and often, those which require multiple demonstrations of an effect across different methods and contexts, again increasing generality. While these criteria all have some merit, a less valid shortcut to the appearance of generality plays on implicit assumptions about people who are seen as normal and usual in research. Just as social psychologists ask Americans to confront the kind of implicit bias that sees "default Americans" as White (e.g., Devos & Banaji, 2005), so the generalizability challenge asks researchers to confront the possibility that they might implicitly see the "default sample" as Western, US American, and/or majority White, with any other characteristics standing in need of special explanation.

## Uptake

Steps to diversify the participant base of psychology over the past ten years, although limited, are worth reviewing. As Thalmeyer et al. (2021) document, the proportion of US-based samples in six high-prestige psychology journals decreased by about 10% since Arnett's (2008) review, but the slack was mostly picked up by researchers in other English-speaking countries and Western Europe. Newson et al. (2021) also found a statistically significant decrease of Western-based studies in mainstream psychology journals over a decade, but only from 95% to 92%, with the difference made up largely by an increase in Asian-based scholarship. The slow pace and scope of progress is not surprising, given that few concrete measures have been taken to promote research on samples from outside the United States and the West more generally (one exception being Kitayama's, 2017, call for more cross-cultural replications in the Attitudes and Social Cognition section of *JPSP*).

Sample composition in social psychology has also seen shifts in the past decade driven by the greater availability of crowdsourced online participants. By 2015, nearly 50% of empirical studies reported in major social psychology journals were using online data collection, with the majority of these being Amazon Mechanical Turk and most of the rest still being university populations (C. A. Anderson et al., 2019). In some ways, this shift relieves earlier worries of a science based almost exclusively on studies with student participants (e.g., Sears, 1986). However, the new samples from Amazon Mechanical Turk, Prolific, and the like are largely Western in origin and not representative even of their national populations, especially in the sheer number of studies a typical worker has completed (e.g., Hargittai & Shaw, 2020; Stewart et al., 2016; Weigold & Weigold, 2022). While some social psychology studies find that convenience samples show similar effects to more general U. S. population samples (e.g., in question framing effects; Mullinix et al., 2015), others show differences, usually finding lower effect sizes in the general population (e.g., in social influence and heuristic effects; Yeager et al., 2019).

External validity concerns are routinely addressed in the Limitations sections of published articles, as B. Clarke et al. (2023) show, but this repeated act of apology does not necessarily translate into a will to improve external validity. Instead, in Clarke et al.'s estimation, it signals "suboptimal practices that a field is willing to acknowledge, perhaps because they are justified in the current context, but not ideal" (p. 896). Some initiatives, all the same, have urged more explicit and pointed

acknowledgements of sample-based limitations. A prominent example is the Constraints on Generality statement, a formal statement of sample- and method-based limitations which has been proposed as a part of the research article's Discussion section (Simons et al., 2017). Constraints on Generality have been implemented in a number of journals, together with like-minded initiatives (e.g., Kitayama, 2017; Lindsay, 2019; the positionality statements covering researchers' perspectives advocated by Roberts et al., 2020). For skeptics who question the visibility of statements included near the end of the paper, the Journal Article Reporting Standards (JARS; Appelbaum et al., 2018), endorsed by the APA and included in the 7th Edition of their Publication Manual, say that an article's title and abstract must contain information on participant source and characteristics. However, it is safe to say that no social psychology journals, APA or otherwise, seem to be enforcing this standard, which would surely lead to an embarrassing preponderance of "in college students" and "in online workers" among tables of contents.

Initiatives such as the Psychological Science Accelerator (Moshontz et al., 2018), already mentioned as a way to enhance the statistical power of research, also diversify the participant and researcher base of psychology (for an overview of distributed research efforts in other fields see Coles et al., 2022). The ability of distributed projects to speak for humanity at large increases when they include Global South samples, in addition to samples from the Western world and East Asian industrialized nations (IJzerman et al., 2021; Silan et al., 2021). As an additional development of note, the ManyLabs Africa project aims to recruit worldwide samples to test ideas and measures developed by African psychologists (Adetula et al., 2022). This reversal of the usual flow of research ideas from center to periphery deserves attention and emulation.

## Costs And Limitations

"Samples of convenience" are called that for a reason. Increasing generalizability by moving outside the mainstream methods, settings, and populations of social psychology presents great costs for the researcher. These costs are not always justified by greater recognition for doing more difficult studies, in a world that counts rather than reads publications when evaluating researchers. What's more, addressing generality concerns in research can mean falling afoul of evidence concerns. The movement toward larger sample sizes, if strictly enforced, can discourage work with harder-to-recruit samples and harder-to-run behavioral methods. The movement toward more open data might inhibit work using samples which present greater confidentiality concerns (Chauvette et al., 2019). Rigid requirements for preregistration, too, might inhibit work among settings and populations for which exploratory and descriptive work are more appropriate.

The solution for all these conflicts is to apply scientific reform with an eye on its ultimate goals, rather than narrowly enforcing specific means to those goals. By being more open-minded about the value of research that adds generalizability and validity, while still being open-eyed about when research truly supports strong inferences, the insights of scientific reform can open up rather than close down the paths available in the research world. In doing so, researchers and research evaluators would do well to keep in mind perspectives on scientific reform that look outside the center of the research establishment (e.g., Lewis Jr., 2021, on connections between central-peripheral and basic-applied relations; Serwadda et al., 2018, on open data; Syed & Kathuwalla, 2021, on cultural psychology and open science).

## New Rules?

As a recognition of the limits of research conducted in a narrow range of settings, we have already seen the adoption of Constraints on Generality or similar statements in some journals (e.g., *Psychological Science; JPSP: Attitudes and Social Cognition*). As I have argued, such steps would be even more effective if carried out in the Title or Abstract of articles, from which many readers draw conclusions without going further. Another minimal step to promote more diverse samples could be taken by raising awareness about sample bias in evaluation. Journals need to explicitly communicate, through policy and reviewer letters, that no human sample should be evaluated as less valid or generalizable than others based on nationality or ethnicity. Although researchers working with minoritized or non-Western samples are sometimes asked in the review process why they don't have, for example, a "White control group" (Chang & Sue, 2005; Markus, 2008), these questions are not appropriate, and less so if researchers using a more usual kind of sample are never asked where their non-White (or non-Western) control group is.

Awareness of sample origin could also be raised by requiring mention of national origin and other demographic characteristics in the Participants section, even if participants come from a more central population. Currently, while the APA Publication Manual (7th edition, 2020) encourages reporting on all manner of participant demographic characteristics, nationality of the sample is not one of them (and yet, oddly, immigration status is). Awareness of US-centrism in other, smaller ways might help reinforce a more inclusive mindset. When participants are described as coming from, for example, "a large Southeastern university," is it assumed that we are talking about Florida or Georgia rather than Guangzhou or Shenzhen? As another example, the US Democratic and Republican parties are sometimes treated as primary ideological categories of interest for our journals (e.g., de Leon et al., 2020; Gelfand, 2022). Would a study of differences between supporters of the Congress and Bharatiya Janata parties in India, which together have more than twice the combined voters of the two main US parties, really hold the same interest to editors and reviewers in a general social psychology journal?

These slight but welcome adjustments to consciousness in publishing would be comparable to leaving our door open. But making sure underrepresented communities arrive at the door in the first place is a complex and difficult task, given the inequalities built into the global research ecosystem. Professional societies based in central regions can play their part by openly acknowledging issues with inclusion and diversity, and offering solutions within their capacity. One example is the Society for Improvement of Psychological Science (SIPS), whose Global Engagement Task Force has called for stronger links with organizations outside the West; consideration of financial and border barriers to conference attendance through virtual participation, sliding fee scales, and diverse locations; and listening to members outside the West through surveys (Steltenpohl et al., 2019).

Faced with these challenges to generalizability, numerous authors over the decades have called for a reassessment of nomothetic science, the usual approach which seeks to test and apply general laws to human behavior. They have proposed that social phenomena might be more sensitively captured by idiographic and critical approaches, which situate and explain individual phenomena in the manner of historiography or ethnography (Gergen, 1973; Langdridge, 2008; Sullivan, 2020). These approaches often reject quantitative research, with its emphasis on establishing parameters of central tendency across many people and settings, in favor of qualitative and descriptive methods (Marecek et al., 1997). Greater acceptance of idiographic research goals, with appropriate cautions against overgeneralizing, would go hand-in-hand with the greater acceptance of descriptive and exploratory research, as suggested in the theory section above. Idiographic methods might be especially appropriate for gaining knowledge of populations whose limited numbers and

accessibility make them difficult to study using the new standards for quantitative inferential research. Inferences from small or limited samples can also be useful when ideas can be disproven, or qualified, via just one disconfirming instance. However, this mode of argument may be better developed in other disciplines than social psychology, such as in political science (Coppedge, 1999; Dion, 1998).

# CONCLUSION

My analysis of methodological challenges in the past decade has identified the failure of several "shortcuts" in research practice and reporting, encouraged by heuristics for judging the quality of published research. It has also reviewed the ongoing process of accepting new standards through new practices and reexamined priorities. The greatest changes have occurred in our ideas of strong evidence. It is no longer generally accepted that all you need to establish the truth of a hypothesis is a series of uniformly significant p-values. Now, questions must be asked about the size of the effect, the quality of methodology (in particular, statistical power), and the completeness of research reporting that led to positive results. A wider view of a finding's context and quality also would ideally lead to a greater acceptance of published negative findings, both within multi-study articles and as conclusions in and of themselves.

Change along all these dimensions has not been total, but there is momentum. Most major journals in social psychology have published editorials and instituted policies acknowledging the need for reform in various ways. Open data and materials, openings for replication studies, and more complete statistical reporting, all are rapidly becoming the norm. Expectations for statistical power in individual studies are increasing, although exact criteria are likely to remain contentious while the field works out what kind of effect sizes are reasonable to expect in our research. Many authors and some journals are taking further steps to establish the robustness of their evidence base through preregistration, disclosure statements, and Registered Reports. And even if journal editors may still not feel comfortable in publishing all null results that come their way, there is greater awareness that suppressing null findings just for being null is wrong, and that traditional excuses for bypassing the APA Publication Manual's long-standing admonitions against this practice need to be re-examined.

In carrying out changes, we must keep in mind that social psychologists are attracted to a field that lets them find answers to individual and social questions through a combination of creativity and rigor. Some reactions against change have painted grim pictures of a scientific dystopia in which reformers end up effectively prohibiting certain kinds of research, whether it is creative exploration that defies *a priori* specification, or ways of knowing that do not fit the normative science of laboratory-based experimentation. As one critic in political science has said of preregistration: "The wonder of discovery is removed from the entire enterprise, rendering the job much closer to that of a mechanic than an artist" (McDermott, 2019, p. 58).

However, changes when properly applied should not suppress creativity at the root. Rather, they should make clear when a given activity is operating in creative mode, and when it has followed the more mechanistic rules of scientific rigor. My essay titled "Science or art?" (Giner-Sorolla, 2012) criticized a process in which the standards of art -- loose in application, but perfectionistic in their expectations about how evidence should look -- were being misapplied to the process of evaluating scientific evidence. Some compromises inevitably must be made if we want our creative ideas to garner the respect that the scientific method claims. That respect is earned precisely from a

willingness to modify or abandon ideas when the data speak against them, using a method of analysis that proceeds according to principle. Generate ideas creatively; but test them rigorously. That is the essence of any science.

To avoid stifling creative and generalizable research, social psychologists who follow and implement these new policies should always keep their ultimate purpose in mind. Well-defined practices and heuristics are helpful in gaining a common understanding of standards. Ideally, they give the sense that appraisal is fair rather than capricious. But they are also prone to ossifying in the hands of gatekeepers, and being exploited by clever gate-leapers, so they must be watched carefully. Keep in mind that there is some redundancy in practices. For example, preregistration's goals overlap greatly with full contemporaneous transparency about methods (Rubin, 2020). Also, preregistration and direct replication are less necessary under rigorous development of theory and accompanying validation of methods (Fiedler, 2017). Thus, if one reform proves difficult or inappropriate in a given setting, there are other ways to reach the goal.

We can also realize that tightening at one end goes together with loosening at another. Many of these changes correct the mislabeling of research practices which editors and reviewers originally demanded to fit into an image of positivistic research. The changes ensure the correct application of labels, but it is just as easy to accept both labeled and unlabeled research for what they are, as to reject the unlabeled product. As examples, preregistered and strongly theorized hypotheses should not pre-empt the publication of exploratory and descriptive analyses. Standards of complete disclosure should go together with relaxed expectations for perfect-looking results. Different ways of knowing might be accepted, especially in places where positivistic science comes up short. My suspicion, however, is that the 80%-plus rejection rates that are currently taken as a mark of publishing quality make it difficult to loosen standards on a large scale. To open the definition of good science while keeping the system as it is, editors would need to work out what the top 20% of descriptive studies, replication studies, exploratory studies, or qualitative studies would look like. This is not an impossible task. But before bending over backwards to keep those 80% rejection rates, we should also consider whether they are more helpful than harmful to scientific communication and career-building.

Ultimately, reforms to the way social psychology is conducted are worthwhile for one final reason: our ethical and moral image in the eyes of people outside social psychology. The common-sense understanding of science includes expectations about integrity. Laypeople think that scientists should report all their evidence, whether or not findings support the idea being tested. Indeed, they see selective reporting as a punishable offense, sometimes judging it as harshly as actual data fabrication (Bottesini et al., 2022; Pickett & Roche, 2018). Laypeople also expect that that scientific findings should be replicable (Fetterman & Sassenberg, 2015; Mountcastle -Shah et al., 2003), and are more likely to trust scientific results if their data and materials have been made open (Schneider et al., 2020). The kind of justifications sometimes heard in defense of questionable practices (e.g., Washburn et al., 2018) -- that journal editors demand them, that suppressing failed results is done to benefit the reader, that whipping datasets into shape for open sharing is too much work -- are unlikely to make a dent in lay perceptions. Common-sense expectations of integrity carry ethical weight. Morality in the present era of psychological research does not inhere in judging specific tools as right or wrong, or in belittling researchers for their past practices. Rather, what is being developed is a system of professional ethics in research and reporting, fit to give robust answers to the serious social issues that empirical social psychologists have always been driven to shed light upon.

# REFERENCES


Abele-Brehm, A. E., Gollwitzer, M., Steinberg, U., & Schönbrodt, F. D. (2019). Attitudes toward Open Science and public data sharing: A survey among members of the German Psychological Society. *Social Psychology*, *50*, 252–260. https://doi.org/10.1027/1864-9335/a000384

Abelson, R. P. (1995). *Statistics as principled argument*. Psychology Press.

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. In *Advances in Methods and Practices in Psychological Science* (pp. 357–366). https://doi.org/10.1177/2515245918773742

Adair, J. G. (2006). Creating indigenous psychologies. In U. Kim, K.-S. Yang, & K.-K. Hwang (Eds.), *Indigenous and cultural psychology* (pp. 467–485). Springer. https://doi.org/10.1007/0-387-28662-4_21

Adams, G., Dobles, I., Gómez, L. H., Kurtiş, T., & Molina, L. E. (2015). Decolonizing psychological science: Introduction to the special thematic section. *Journal of Social and Political Psychology*, *3*, 213-238 10 5964 3 1 564.

Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from—Not just to—Africa. *Nature Reviews Psychology*, *1*, 370–371. https://doi.org/10.1038/s44159-022-00070-y

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Aldhous, P. (2011). Journal rejects studies contradicting precognition. *New Scientist*, 2514 5.

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*, 146. https://doi.org/10.1037/amp0000258

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*, 842–850. https://doi.org/10.1177/0146167218798821

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods*, *25*, 596–609. https://doi.org/10.1037/met0000248

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*, 1547–1562. https://doi.org/10.1177/0956797617723724

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104–159. https://doi.org/10.1016/j.jesp.2021.104159

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*, 3–25. https://doi.org/10.1037/amp0000191

Arnett, J. J. (2009). The neglected 95%, a challenge to psychology's philosophy of science. *American Psychologist*, *64*, 571–574. https://doi.org/10.1037/a0016723

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*, 527–546. https://doi.org/10.1037/met0000365

Association, A. P. (1994). *Publication manual of the American Psychological Association* (4th ed.). American Psychological Association.

Association, A. P. (2003). *Publication manual of the American Psychological Association* (5th ed.). American Psychological Association.

Association, A. P. (2010). *Preparing manuscripts for publication in psychology journals: A guide for new authors*. https://www.apa.org/pubs/authors/new-author-guide.pdf

Association, A. P. (2020). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.

Association, A. P. (2021, August). *Transparency and Openness Promotion (TOP) guidelines at APA journals*. https://www.apa.org/pubs/journals/resources/top-guidelines-journal-flyer.pdf

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts. *One Data Set" Project. Socius*, *7*, 23780231211024421. https://doi.org/10.1177/23780231211024421

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, *8*, 1–7. https://doi.org/10.1186/1741-7015-8-17

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, *7*, 1247. https://doi.org/10.3389/fpsyg.2016.01247

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bakker, M., Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., Veldkamp, C. L., Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*, 3000937. https://doi.org/10.1371/journal.pbio.3000937

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Bartlett, N. (2014, June 23). *Replication crisis in psychology research turns ugly and odd*. The Chronicle of Higher Education. https://www.chronicle.com/article/replicationcrisis-

Bastian, H. (2014). A stronger post-publication culture is needed for better science. *PLOS Medicine*, *11*, 1001772. https://doi.org/10.1371/journal.pmed.1001772

Bauer, M. A., Wilkie, J. E., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism: Situational materialism undermines personal and social well-being. *Psychological Science*, *23*, 517–523. https://doi.org/10.1177/0956797611429579

Baumeister, R. F., Bratslavsky, M., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*, 1252–1265. https://doi.org/10.1037//0022-3514.74.5.1252

Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation*. Academic Press.

Baumeister, R. F., Muraven, M., & Tice, D. M. (2000). Ego depletion: A resource model of volition, self-regulation, and controlled processing. *Social Cognition*, *18*, 130–150. https://doi.org/10.1521/soco.2000.18.2.130

Baumeister, R. F., Tice, D. M., & Bushman, B. J. (2023). A review of multisite replication projects in social psychology: Is it viable to sustain any confidence in social psychology's knowledge base? *Perspectives on Psychological Science*, *18*(4), 912–935. https://doi.org/10.1177/17456916221121815

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. https://doi.org/10.3758/s13428-011-0081-0

Bem, D. J. (2003). *Writing the empirical journal article* (2nd ed.). American Psychological Association. https://doi.org/10.4324/9781315808314-10

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berkman, E. T., & Wilson, S. M. (2021). So useful as a good theory? The practicality crisis in (social) psychological theory. *Perspectives on Psychological Science*, *16*, 864–874. https://doi.org/10.1177/1745691620969650

Bishop, D. B. (2014, October 16). *Replication is not about making or breaking careers*. https://blogs.lse.ac.uk/impactofsocialsciences/2014/10/16/replication-andreputation-

Bless, H., & Burger, A. M. (2016). A closer look at social psychologists' silver bullet: Inevitable and evitable side effects of the experimental approach. *Perspectives on Psychological Science*, *11*, 296–308. https://doi.org/10.1177/1745691615621278

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLOS Medicine*, *11*, 1001607. https://doi.org/10.1371/journal.pmed.1001607

Borgatta, E. F., & Bohrnstedt, G. W. (1974). Some limitations on generalizability from social psychological experiments. *Sociological Methods and Research*, *3*, 111–120. https://doi.org/10.1177/004912417400300105

Bornmann, L., & Marx, W. (2012). The Anna Karenina principle: A way of thinking about success in science. *Journal of the American Society for Information Science and Technology*, *63*, 2037–2051. https://doi.org/10.1002/asi.22661

Borsboom, D., Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047

Bosnjak, M., Fiebach, C. J., Mellor, D., Mueller, S., O'Connor, D. B., Oswald, F. L., & Sokol, R. I. (2022). A template for preregistration of quantitative research in psychology: Report of the joint psychological societies preregistration task force. *American Psychologist*, *77*, 602–615. https://doi.org/10.1037/amp0000879

Bottesini, J. G., Rhemtulla, M., & Vazire, S. (2022). What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society Open Science*, *9*, 200048. https://doi.org/10.1098/rsos.200048

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., & Rieck, J. R. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88.

Bou Zeineddine, F., Saab, R., Lášticová, B., Kende, A., & Ayanian, A. H. (2022). Some uninteresting data from a faraway country": Inequity and coloniality in international social psychological publications. *Journal of Social Issues*, *78*, 320–345. https://doi.org/10.1111/josi.12481

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication. *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Bressan, P. (2019). Confounds in 'failed' replications. *Frontiers in Psychology*, *10*, 1884. https://doi.org/10.3389/fpsyg.2019.01884

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*, 980–989.

https://doi.org/10.1038/s41562-021-01143-3

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. https://doi.org/10.1177/1745691610393980

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer. *Journal of Personality and Social Psychology*, *98*(4), 550–558. https://doi.org/10.1037/a0018933

Cairo, A. H., Green, J. D., Forsyth, D. R., Behler, A. M. C., & Raldiris, T. L. (2020). Gray (literature) matters: Evidence of selective hypothesis reporting in social psychological research. *Personality and Social Psychology Bulletin*, *46*, 1344–1362. https://doi.org/10.1177/0146167220903896

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*, 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. https://doi.org/10.1037/h0046016

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233–239. https://doi.org/10.1177/2515245919858072

Castro Torres, A. F., & Alburez-Gutierrez, D. (2022). North and South: Naming practices and the hidden dimension of global disparities in knowledge production. *Proceedings of the National Academy of Sciences*, *119*, 2119373119. https://doi.org/10.1073/pnas.2119373119

Ceci, S. J., & Walker, E. (1983). Private archives and public needs. *American Psychologist*, *38*, 414–423. https://doi.org/10.1037/0003-066X.38.4.414

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

Cesario, J. (2022). What can experimental studies of bias tell us about real-world group disparities? *Behavioral and Brain Sciences*, *45, e66*, 1–71. https://doi.org/10.1017/S0140525X21000017

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. D., & Mellor, D. T. (2018). Protocol transparency is vital for registered reports. *Nature Human Behaviour*, *2*, 791–792. https://doi.org/10.1038/s41562-018-0449-6

Chauvette, A., Schick-Makaroff, K., & Molzahn, A. E. (2019). Open data in qualitative research. *International Journal of Qualitative Methods*, *18*, 1609406918823863. https://doi.org/10.1177/1609406918823863

Cheon, B. K., Melani, I., & Hong, Y. Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study

samples. *Social Psychological and Personality Science*, *11*, 928–937. https://doi.org/10.1177/1948550620927269

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*, 377–395. https://doi.org/10.1177/1745691620950684

Chuard, P. J., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse p-hacking or selective reporting. *PLOS Biology*, *17*, 3000127. https://doi.org/10.1371/journal.pbio.3000127

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science*, *4*, 5–6. https://doi.org/10.1111/j.1745-6924.2009.01091.x

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*, 211037. https://doi.org/10.1098/rsos.211037

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022). Keep your enemies close: Adversarial collaborations will improve behavioral science. *Journal of Applied Research in Memory and Cognition*, *11*, 1–18. https://doi.org/10.1037/mac0000004

Clarke, B., Schiavone, S., & Vazire, S. (2023). What limitations are reported in short articles in social and personality psychology? *Journal of Personality and Social Psychology*, *125*, 874–901. https://doi.org/10.1037/pspp0000458

Cohen, J. (1969). *Statistical power analysis for the biomechanical sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin*, *112*, 1155–1159. https://doi.org/10.1037/0033-2909.112.1.155

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*, 505–507. https://doi.org/10.1038/d41586-022-00150-2

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 4716. https://doi.org/10.1126/science.aac4716

Collective, D. P. E. (2021). General psychology otherwise: A decolonial articulation. *Review of General Psychology*, *25*, 339–353. https://doi.org/10.1177/10892680211048177

Collins, H. (1985). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Cook, T. D., & Groom, C. (2004). The methodological assumptions of social psychology: The mutual dependence of substantive theory and method choice. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage Handbook of Methods in Social Psychology* (pp. 19–44). Sage Publications, Inc. https://doi.org/10.4135/9781412976190.n2

Coppedge, M. (1999). Thickening thin concepts and theories: Combining large N and small in comparative politics. *Comparative Politics*, *31*(4), 465–476. https://doi.org/10.2307/422240

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the. *Journal of Applied Psychology. Journal of Applied Psychology*, *105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, *17*, 136–137. https://doi.org/10.1037/0735-7028.17.2.136

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

Crutzen, R., & Peters, G. J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*(3), 242–247. https://doi.org/10.1080/17437199.2015.1124240

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., & Brown, N. J. (2023). What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of Psychological Science. *Psychological Science*, *34*(4), 512–522. https://doi.org/10.1177/09567976221140828

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. https://doi.org/10.4324/9780203807002

DeJesus, J. M., Callanan, M. A., Solis, G., & A, G. S. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, *116*, 18370–18377. https://doi.org/10.1073/pnas.1817706116

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*, 2515245920954925. https://doi.org/10.1177/2515245920954925

Derksen, M., & Field, S. (2022). The tone debate: Knowledge, self, and social order. *Review of General Psychology*, *26*, 172–183. https://doi.org/10.1177/10892680211015636

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805. https://doi.org/10.1098/rsos.200805

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447–466. https://doi.org/10.1037/0022-3514.88.3.447

Dion, D. (1998). Evidence and inference in the comparative case study. *Comparative Politics*, *30*, 127–146. https://doi.org/10.2307/422284

Do people embrace praise even when they feel unworthy? A review of critical tests of self-enhancement versus self-verification. (2010). *Personality and Social Psychology Review*, *14*(3), 263–280. https://doi.org/10.1177/1088868310365876

Doliński, D. (2018). Is psychology still a science of behaviour? *Social Psychological Bulletin*, *13*, 25025. https://doi.org/10.5964/spb.v13i2.25025

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind. *PLOS ONE*, *7*, 29081. https://doi.org/10.1371/journal.pone.0029081

Dutilh, G., Sarafoglou, A., & Wagenmakers, E. J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*(23), 5745–5772. https://doi.org/10.1007/s11229-019-02456-7

Dvorak, R. D., & Simons, J. S. (2009). Moderation of resource depletion in the self-control strength model: Differing effects of two modes of self-control. *Personality and Social Psychology Bulletin*, *35*, 572–583. https://doi.org/10.1177/0146167208330855

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., & 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

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. https://doi.org/10.1177/0956797613512465

Ejelöv, E., & Luke, T. J. (2020). Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, *87*, 103937. https://doi.org/10.1016/j.jesp.2019.103937

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

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, *30*, 967–976. https://doi.org/10.1037/0003-066X.30.10.967

Engber, D. (2017). Daryl Bem proved ESP is real: Which means science is broken. *Slate*. https://slate.com/

Eriksson, K., & Simpson, B. (2013). Editorial decisions may perpetuate belief in invalid research findings. *PloS One*, *8*(9), 73364. https://doi.org/10.1371/journal.pone.0073364

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on The Reproducibility Project: Psychology. *PLOS One*, *11*, 0149794. https://doi.org/10.1371/journal.pone.0149794

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. https://doi.org/10.1016/j.jesp.2015.07.009

Fanelli, D. (2010). Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*, 10068. https://doi.org/10.1371/journal.pone.0010068

Fayant, M. P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology*, *30*, 125–130. https://doi.org/10.5334/irsp.102

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing. *PLOS ONE*, *10*, 0118053. https://doi.org/10.1371/journal.pone.0118053

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538. https://doi.org/10.1037/a0015808

Fetterman, A. K., & Sassenberg, K. (2015). The reputational consequences of failed replications and wrongness admission among scientists. *PLOS ONE*, *10*, 0143723. https://doi.org/10.1371/journal.pone.0143723

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Cognitive Science*, *12*, 46–61. https://doi.org/10.1177/1745691616654458

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*, 433–438. https://doi.org/10.1177/1745691617745651

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests-An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*, 816–826. https://doi.org/10.1177/1745691620970602

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231–1236. https://doi.org/10.1016/j.jesp.2011.05.007

Fife, D. (2020). The eight steps of data analysis: A graphical framework to promote sound statistical analysis. *Perspectives on Psychological Science*, *15*(4), 1054–1075. https://doi.org/10.1177/1745691620917333

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*, 244–253. https://doi.org/10.1037/pspi0000075

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, *8*, 053. https://doi.org/10.1093/gigascience/giz053

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science*, *8*, 370–378. https://doi.org/10.1177/1948550617693063

Forestier, C., Chanaleilles, M., Boisgontier, M. P., & Chalabaev, A. (2022). From ego depletion to self-control fatigue: A review of criticisms along with new perspectives for the investigation and replication of a multicomponent phenomenon. *Motivation Science*, *8*, 19–32. https://doi.org/10.1037/mot0000262

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*, 109019. https://doi.org/10.1371/journal.pone.0109019

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991. https://doi.org/10.3758/s13423-012-0322-y

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess success for psychology articles in the journal Science. *PLOS ONE*, *9*, 114255. https://doi.org/10.1371/journal.pone.0114255

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*, 8–12. https://doi.org/10.1177/1948550615598377

Frank, M. (2015, August 31). *The slower, harder ways to increase reproducibility*. http://babieslearninglanguage.blogspot.com/2015/08/the-slower-harder-ways-toincrease.

Friese, M., & Frankenbach, J. (2020). P-Hacking and publication bias interact to distort metaanalytic effect size estimates. *Psychological Methods*, *25*, 456–471. https://doi.org/10.1037/met0000246

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real? An analysis of arguments. *Personality and Social Psychology Review*, *23*, 107–131. https://doi.org/10.1177/1088868318762183

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology*, *23*, 98–122. https://doi.org/10.1177/0959354312436870

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18. https://doi.org/10.1037/a0024338

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*, 156–168. https://doi.org/10.1177/2515245919847202

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*, 325–336. https://doi.org/10.1037/0022-3514.92.2.325

Galak, J., LeBoeuf, R. A., Nelson, L., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*, 933–948. https://doi.org/10.1037/a0029709

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., & Dimant, E. (2022). Persuading Republicans and Democrats to comply with mask wearing: An intervention tournament. *Journal of Experimental Social Psychology*, *101*, 104299. https://doi.org/10.1016/j.jesp.2022.104299

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis-a ‘ garden of forking paths’–explains why many statistically significant comparisons don’t hold up. *American Scientist*, *102*, 460–466. https://doi.org/10.1511/2014.111.460

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*, 309–320. https://doi.org/10.1037/h0034436

Ghai, S. (2021). It’s time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour*, *5*(8), 971–972. https://doi.org/10.1038/s41562-021-01175-9

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*(2), 195–204. https://doi.org/10.1177/0959354398082006

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (pp. 391–408). Sage Publications.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511720482

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on ‘Estimating the reproducibility of psychological science’. *Science*, *351*, 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*, 562–571. https://doi.org/10.1177/1745691612457576

Giner-Sorolla, R. (2016). Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology*, *65*, 1–6. https://doi.org/10.1016/j.jesp.2016.01.010

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*, 1–38. https://doi.org/10.1080/10463283.2018.1542902

Giner-Sorolla, R., Amodio, D. M., & Kleef, G. (2018). Three strong moves to improve research and replications alike. *Behavioral and Brain Sciences*, *41*, 140. https://doi.org/10.1017/S0140525X18000651

Giner-Sorolla, R., Montoya, A. K., Reifman, A., Carpenter, T., Lewis, N. A., Aberson, C. L., Bostyn, D. H., Conrique, B. G., Ng, B. W., Schoemann, A. M., & Soderberg, C. (2024). Power to detect what? Considerations for planning and evaluating sample size. *Personality and Social Psychology Review*, *28*(3), 276–301. https://doi.org/10.1177/10888683241228328

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535–549. https://doi.org/10.1111/spc3.12267

Goldfried, M. R. (1959). One-tailed tests and ‘unexpected’ results. *Psychological Review*, *66*, 79–80. https://doi.org/10.1037/h0038521

Goldin-Meadow, S. (2016). Why preregistration makes me nervous. *APS Observer*, *29*.

Gollwitzer, M., & Schwabe, J. (2022). Context dependency as a predictor of replicability. *Review of General Psychology*, *26*, 241–249. https://doi.org/10.1177/10892680211015635

Goodhart, C. (1984). Problems of monetary management: The UK experience. In *Monetary theory and practice* (pp. 91–121). Palgrave. https://doi.org/10.1007/978-1-349-17295-5_4

Goodman, S. N. (2019). Why is getting rid of p-values so hard? Musings on science and statistics. *The American Statistician*, *73*(sup1), 26–30. https://doi.org/10.1080/00031305.2018.1558111

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*, 205–215. https://doi.org/10.1177/1745691620984483

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*, 803–815. https://doi.org/10.1177/1745691620966794

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science*, *12*, 731–741. https://doi.org/10.1177/1745691617691949

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

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*, 99–108. https://doi.org/10.1177/1745691611434210

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*, 216–229. https://doi.org/10.1037//0033-295X.93.2.216

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated. *Psychophysiology*, *33*, 175–183. https://doi.org/10.1111/j.1469-8986.1996.tb02121.x

Greenwood, J. D. (2000). Individualism and the social in early American social psychology. *Journal of the History of the Behavioral Sciences*, *36*, 443–455. https://doi.org/10.1002/1520-6696(200023)36:4

Gregg, A. P., & Sedikides, C. (2004). Is social psychological research really so negatively biased? *Behavioral and Brain Sciences*, *27*(3), 340–341. https://doi.org/10.1017/S0140525X04340082

Groot, A. D. (1956). The meaning of 'significance' for different types of research. *Acta Psychologica*, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*, 1243–1255. https://doi.org/10.1177/1745691620921521

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in Psychological Science. *Perspectives on Psychological Science*, *16*, 789–802. https://doi.org/10.1177/1745691620970585

Haig, B. D. (2022). Understanding replication in a way that is true to science. *Review of General Psychology*, *26*, 224–240. https://doi.org/10.1177/10892680211046514

Hanel, P. H., Maio, G. R., & Manstead, A. S. (2019). A new way to look at the data: Similarities between groups of people are large and important. *Journal of Personality and Social Psychology*, *116*, 541–562. https://doi.org/10.1037/pspi0000154

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., & 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), 201494. https://doi.org/10.1098/rsos.201494

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017. *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hargittai, E., & Shaw, A. (2020). Comparing internet experiences and prosociality in Amazon Mechanical Turk and population-based survey samples. *Socius*, *6*, 2378023119889834. https://doi.org/10.1177/2378023119889834

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (pp. 3–24). American Psychological Association. https://doi.org/10.1037/0000135-001

Hastie, R., & Stasser, G. (2000). Computer simulation methods for social psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 85–114). Cambridge University Press.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 998. https://doi.org/10.3389/fpsyg.2018.00998

Haven, T. L., & Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244. https://doi.org/10.1080/08989621.2019.1580147

Hegarty, P., & Bruckmüller, S. (2013). Asymmetric explanations of group differences: Experimental evidence of Foucault's disciplinary power. *Social and Personality Psychology Compass*, *7*(3), 176–186. https://doi.org/10.1111/spc3.12017

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, *5*, 41–49.

Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis= trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science*, *29*(3), 270–288. https://doi.org/10.1177/0963662520902383

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The WEIRDest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Heo, M., Kim, N., & Faith, M. S. (2015). Statistical power as a function of Cronbach alpha of instrument questionnaire items. *BMC Medical Research Methodology*, *15*, 86. https://doi.org/10.1186/s12874-015-0070-6

Hodson, G. (2021). Construct jangle or construct mangle? Thinking straight about (nonredundant) psychological constructs. *Journal of Theoretical Social Psychology*, *5*, 576–590. https://doi.org/10.1002/jts5.120

Hogg, M. A., & Williams, K. D. (2000). From I to We: Social identity and the collective self. *Group Dynamics: Theory, Research, and Practice*, *4*, 81–97. https://doi.org/10.1037/1089-2699.4.1.81

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. https://doi.org/10.1177/2515245920919667

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., & Nilsonne, G. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, *12*, 1–47. https://doi.org/10.31234/osf.io/pbfye

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*, 70–85. https://doi.org/10.1177/2515245917751886

Hruschka, D. J., Medin, D. L., Rogoff, B., & Henrich, J. (2018). Pressing questions in the study of psychological and behavioral diversity. *Proceedings of the National Academy of Sciences*, *115*(45), 11366–11368. https://doi.org/10.1073/pnas.1814733115

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *Journal of Applied Statistics*, *38*, 2617–2626. https://doi.org/10.1080/02664763.2011.567245

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 166–184. https://doi.org/10.1177/2515245919882903

IJzerman, H., Dutra, N., Silan, M., Adetula, A., Brown, D. M. B., & Forscher, P. (2021). Psychological science needs the entire globe. *APS Observer*, *34*.

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., Ivory, J. D., & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behaviour*, *4*, 1092–1094. https://doi.org/10.1038/s41562-020-00990-w

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences*, *113*(34), 4933–4934. https://doi.org/10.1073/pnas.1608676113

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*, 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*, *50*(5–6), 370–378. https://doi.org/10.1027/1864-9335/a000398

Inzlicht, M., & Schmeichel, B. J. (2012). what is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, *7*, 450–463. https://doi.org/10.1177/1745691612454134

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*, 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Ioannou, A., Tussyadiah, I., Miller, G., Li, S., & Weick, M. (2021). Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis. *PLOS ONE*, *16*, 0256822. https://doi.org/10.1371/journal.pone.0256822

Isager, P. M., Aert, R., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., & Lakens, D. (2023). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*, *28*, 432–451. https://doi.org/10.1037/met0000438

Iverson, G. J., Lee, M. D., & Wagenmakers, E. J. (2009). $p_{rep}$ misestimates the probability of replication. *Psychonomic Bulletin & Review*, *16*, 424–429. https://doi.org/10.3758/PBR.16.2.424

Jamieson, K. H. (2018). Crisis or self-correction: Rethinking media narratives about the wellbeing of science. *Proceedings of the National Academy of Sciences*, *115*, 2620–2627. https://doi.org/10.1073/pnas.1708276114

Jeon, M., & Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*, 340–360. https://doi.org/10.1037/met0000140

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion-Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science*, *21*, 1686–1693. https://doi.org/10.1177/0956797610384745

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*, 263–274. https://doi.org/10.1007/s10676-014-9351-8

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*, *1*(1–3), 1–7. https://doi.org/10.1080/23743603.2015.1070611

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., & Sirota, M. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169.

Jordan, C. H., & Zanna, M. P. (1999). How to read a journal article in social psychology. In R. F. Baumeister (Ed.), *The Self in Social Psychology* (pp. 461–470). Psychology Press.

Kahalon, R., Klein, V., Ksenofontov, I., Ullrich, J., & Wright, S. C. (2022). Mentioning the sample's country in the article's title leads to bias in research evaluation. *Social Psychological and Personality Science*, *13*, 352–361. https://doi.org/10.1177/19485506211024036

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*, 723–730. https://doi.org/10.1037/0003-066X.58.9.723

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131–1142. https://doi.org/10.1177/0146167208331253

Kelley, E. L. (1927). *Interpretation of educational measurements*. World.

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*, 578–589. https://doi.org/10.1037/met0000209

Kidd, R. F. (1976). Manipulation checks: Advantage or disadvantage? *Representative Research in Social Psychology*, *7*, 160–165.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., & Falkenberg, L.-S. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*, 1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr, Alper, S., & Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, *24*(3), 326–338. https://doi.org/10.1177/0959354314529616

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. American Psychological Association. https://doi.org/10.1037/10693-000

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology*, *65*, 1132. https://doi.org/10.1037/0022-3514.65.6.1132

Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology*, *8*, 244–259. https://doi.org/10.1177/147470491000800208

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*, 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*, 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., & Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*, 259–269.

https://doi.org/10.1177/2515245918770963

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*, 518–536. https://doi.org/10.1177/1745691615583128

Landy, J. F., Jia, M. (L.), 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*, 451–479. https://doi.org/10.1037/bul0000220

Langdridge, D. (2008). Phenomenology and critical social psychology: Directions and debates in theory and research. *Social and Personality Psychology Compass*, *2*(3), 1126–1142. https://doi.org/10.1111/j.1751-9004.2008.00114.x

Lange, F. (2020). Are difficult-to-study populations too difficult to study in a reliable way? Lessons learned from meta-analyses in clinical neuropsychology. *European Psychologist*, *25*(1), 41–50. https://doi.org/10.1027/1016-9040/a000384

LeBel, E. P. (2015). A new replication norm for psychology. *Collabra*, *1: 4*, 1–13. https://doi.org/10.1525/collabra.23

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's. *Review of General Psychology*, *15*, 371–379. https://doi.org/10.1037/a0025172

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*, 254–261. https://doi.org/10.1037/pspi0000106

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*, 230–243. https://doi.org/10.1037/pspi0000049

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences*, *115*(45), 10516–10517. https://doi.org/10.1073/pnas.1812592115

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in Psychological Science. *Perspectives on Psychological Science*, *7*, 60–66. https://doi.org/10.1177/1745691611427304

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003. *Psychological Review*, *112*, 662–668. https://doi.org/10.1037/0033-295X.112.3.662

Legate, N., Ngyuen, T. V., Weinstein, N., Moller, A., Legault, L., Vally, Z., & Ogbonnaya, C. E. (2022). A global experiment on motivating social distancing during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, *119*(22), 2111091119. https://doi.org/10.1073/pnas.2111091119

Lench, H. C., Taylor, A. B., & Bench, S. W. (2014). An alternative approach to analysis of mental states in experimental social cognition research. *Behavioral Research Methods*, *46*, 215–228. https://doi.org/10.3758/s13428-013-0351-0

Leon, R. P., Wingrove, S., & Kay, A. C. (2020). Scientific skepticism and inequality: Political and ideological roots. *Journal of Experimental Social Psychology*, *91*, 104045. https://doi.org/10.1016/j.jesp.2020.104045

Leung, A. K. Y., Kim, S., Polman, E., Ong, L. S., Qiu, L., Goncalo, J. A., & Sanchez-Burks, J. (2012). Embodied metaphors and creative 'acts'. *Psychological Science*, *23*, 502–509. https://doi.org/10.1177/0956797611429801

Levine, M., & Ensom, M. H. (2001). Post hoc power analysis: An idea whose time has passed. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *21*, 405–409. https://doi.org/10.1592/phco.21.5.405.34503

Lewandowsky, S., & Bishop, D. (2016). Research integrity: Don't let transparency damage science. *Nature*, *529*(7587), 459–461. https://doi.org/10.1038/529459a

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*, 1–12. https://doi.org/10.1038/s41467-019-14203-0

Lewin, K. (1951). *Field theory in social sciences*. Harper & Row.

Lewin, M. A. (1977). Kurt Lewin's view of social psychology: The crisis of 1977 and the crisis of 1927. *Personality and Social Psychology Bulletin*, *3*, 159–172. https://doi.org/10.1177/014616727700300203

Lewis, M., Mathur, M. B., VanderWeele, T. J., & Frank, M. C. (2022). The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, *9*, 211499. https://doi.org/10.1098/rsos.211499

Lewis, N. A., Jr. (2021). What counts as good science? How the battle for methodological legitimacy affects public psychology. *The American Psychologist*, *76*, 1323–1333. https://doi.org/10.1037/amp0000870

Li, J., Kim, C. Y., Karp, S. R., & Takooshian, H. (2012). A data-based profile of US social psychology journals: 25 years later. *International Psychology Bulletin*, *16*, 23–29.

Linde, M., Tendeiro, J. N., Selker, R., Wagenmakers, E.-J., & Ravenzwaaij, D. (2021). Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *Psychological Methods*, *28*, 740–755. https://doi.org/10.31234/osf.io/bh8vu

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Lindsay, D. S. (2017). Preregistered direct replications in Psychological Science. *Psychological Science*, *28*, 1191–1192. https://doi.org/10.1177/0956797617718802

Lindsay, D. S. (2019). Swan song editorial. *Psychological Science*, *30*, 1669–1673. https://doi.org/10.1177/0956797619893653

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Lurquin, J. H., & Miyake, A. (2017). Challenges to ego-depletion research go beyond the replication crisis: A need for tackling the conceptual crisis. *Frontiers in Psychology*, *8*, 568. https://doi.org/10.3389/fpsyg.2017.00568

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. https://doi.org/10.1016/j.jesp.2016.09.006

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159. https://doi.org/10.1037/h0026141

Madill, A., & Gough, B. (2008). Qualitative research and its place in Psychological Science. *Psychological Methods*, *13*(3), 254–271. https://doi.org/10.1037/a0013220

Mahajan, N., Martinez, M., Gutierrez, N. L., Diesendruck, G., Banaji, M., & Santos, L. R. (2011). The evolution of intergroup bias: Perceptions and attitudes in rhesus macaques. *Journal of Personality and Social Psychology*, *100*, 387–405. https://doi.org/10.1037/a0022459

Malich, L., & Munafò, M. R. (2022). Replication of crises: Interdisciplinary reflections on the phenomenon of the replication crisis in psychology. *Review of General Psychology*, *26*, 127–130. https://doi.org/10.1177/10892680221077997

Marecek, J., Fine, M., & Kidder, L. (1997). Working between worlds: Qualitative methods and social psychology. *Journal of Social Issues*, *53*, 631–644. https://doi.org/10.1111/j.1540-4560.1997.tb02452.x

Markus, H. R. (2008). Pride, prejudice, and ambivalence: Toward a unified theory of race and ethnicity. *American Psychologist*, *63*(8), 651–670. https://doi.org/10.1037/0003-066X.63.8.651

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, 523. https://doi.org/10.3389/fpsyg.2017.00523

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163. 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*, 487–498. https://doi.org/10.1037/a0039400

Mayo, D. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

McDermott, R. (2022). Breaking free: How preregistration hurts scholars and science. *Politics and the Life Sciences*, *41*, 55–59. https://doi.org/10.1017/pls.2022.4

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289–374. https://doi.org/10.1037/h0060985

McPhetres, J., Albayrak-Aydemir, N., Barbosa Mendes, A., Chow, E. C., Gonzalez-Marquez, P., Loukras, E., & Maus, A. (2021). A decade of theory as reflected in Psychological Science (2009-2019. *PLOS One*, *16*(3), 0247986. https://doi.org/10.1371/journal.pone.0247986

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in metaanalysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749. https://doi.org/10.1177/1745691616662243

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115. https://doi.org/10.1086/288135

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*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meier, B. P., Schnall, S., Schwarz, N., & Bargh, J. A. (2012). Embodiment in social psychology. *Topics in Cognitive Science*, *4*, 705–716. https://doi.org/10.1111/j.1756-8765.2012.01212.x

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology*, *1*, 115–126.

Miller, J., & Ulrich, R. (2022). Optimizing research output: How can psychological research methods be improved. *Annual Review of Psychology*, *73*, 691–718. https://doi.org/10.1146/annurev-psych-020821-094927

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

Montoya, A. K., Krenzer, W. L. D., & Fossum, J. L. (2021). Opening the door to registered reports: Census of journals publishing registered reports (2013-2020. *Collabra*, *7*, 24404. https://doi.org/10.1525/collabra.24404

Mordkoff, J. T. (2019). A simple method for removing bias from a popular measure of standardized effect size: Adjusted partial eta squared. *Advances in Methods and Practices in Psychological Science*, *2*(3), 228–232. https://doi.org/10.1177/2515245919855053

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515. https://doi.org/10.1177/2515245918797607

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

Mountcastle-Shah, E., Tambor, E., Bernhardt, B. A., Geller, G., Karaliukas, R., Rodgers, J. E., & Holtzman, N. A. (2003). Assessing mass media reporting of disease-related genetic discoveries: Development of an instrument and initial findings. *Science Communication*, *24*, 458–478. https://doi.org/10.1177/1075547003024004003

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*, 109–138. https://doi.org/10.1017/XPS.2015.19

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

Nelson, L. D., Gonzalez, F., O'Donnell, M., & Perfecto, H. (2019). Using p-curve to assess evidentiary value from 10 years of published literature. In R. Bagchi, L. Block, & L. Lee (Eds.), *Advances in Consumer Research* (Vol. 47, pp. 212–216). Association for Consumer Research.

Neuliep, J. W. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, *5*, 85–90.

Newson, M., Buhrmester, M., Xygalatas, D., & Whitehouse, H. (2021). Go WILD, Not WEIRD. *Journal for the Cognitive Science of Religion*, *6*(1–2), 80–106. https://doi.org/10.1558/jcsr.38413

Nicolas, G., Bai, X., & Fiske, S. T. (2019). Exploring research-methods blogs in psychology: Who posts what about whom, and with what effect. *Perspectives on Psychological Science*, *14*, 691–704. https://doi.org/10.1177/1745691619835216

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed cancels the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*, 657–664. https://doi.org/10.1037/pspa0000121

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., & Yarkoni, T. (2015). Promoting an open research culture: The TOP guidelines for journals. *Science*, *348*, 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., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*, 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). Reply to Ledgerwood: Predictions without analysis plans are inert. *Proceedings of the National Academy of Sciences*, *115*, 10518–10518. https://doi.org/10.1073/pnas.1816418115

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*, 615–631. https://doi.org/10.1177/1745691612459058

Nuijten, M. B., & Polanin, J. R. (2020). statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, *11*, 574–579. https://doi.org/10.1002/jrsm.1408

O'Donnell, J. M., & G, E. (1979). The crisis of experimentalism in the 1920s. *American Psychologist*, *34*, 289–295. https://doi.org/10.1037/0003-066X.34.4.289

Olsson-Collentine, A., Wicherts, J. M., & Assen, M. A. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*, 922–940. https://doi.org/10.1037/bul0000294

Oransky, I., & Marcus, A. (2016). How researchers lock up their study data with sharing fees. *Stat+*. https://www.statnews.com/2016/09/30/data-sharing-fees/

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*, 539–544. https://doi.org/10.1177/1745691616646366

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

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

Petrescu, M., & Krishen, A. S. (2022). The evolving crisis of the peer-review process. *Journal of Marketing Analytics*, *10*, 185–186. https://doi.org/10.1057/s41270-022-00176-5

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, *31*, 163–176. https://doi.org/10.1002/jcpy.1209

Pickett, J. T., & Roche, S. P. (2018). Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Science and Engineering Ethics*, *24*, 151–171. https://doi.org/10.1007/s11948-017-9886-2

Pittelkow, M. M., Field, S. M., Isager, P. M., Veer, A. E., Anderson, T., Cole, S. N., & Ravenzwaaij, D. (2023). The process of replication target selection in psychology: What to consider? *Royal Society Open Science*, *10*(2), 210586. https://doi.org/10.1098/rsos.210586

PLOS, O. N. E. (2019, December 5). *Data availability*. https://journals.plos.org/plosone/s/data-availability

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson. https://doi.org/10.1063/1.3060577

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*, 1036–1042. https://doi.org/10.1177/0956797616645672

R., A. O., M., A. M. A. L., & Enting, M. (2023). Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231187988

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National*

*Academy of Sciences*, *115*, 11401–11405. https://doi.org/10.1073/pnas.1721165115

Reddy, G., & Amer, A. (2023). Precarious engagements and the politics of knowledge production: Listening to calls for reorienting hegemonic social psychology. *British Journal of Social Psychology*, *62*, 71–94. https://doi.org/10.1111/bjso.12609

Reis, H. T., & Stiller, J. (1992). Publication trends in JPSP: A three-decade review. *Personality and Social Psychology Bulletin*, *18*, 465–472. https://doi.org/10.1177/0146167292184011

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive facilitation of recall' effect. *PLOS ONE*, *7*(3), 33423. https://doi.org/10.1371/journal.pone.0033423

Ritter, F. E., Kim, J. W., Morgan, J. H., & Carlson, R. A. (2012). *Running behavioral studies with human participants: A practical guide*. Sage Publications. https://doi.org/10.4135/9781452270067

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*, 1295–1309. https://doi.org/10.1177/1745691620927709

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. https://doi.org/10.1177/2515245917745629

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to PROCESS! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*, 25152459221095827. https://doi.org/10.1177/25152459221095827

Romero, F. (2020). The division of replication labor. *Philosophy of Science*, *87*, 1014–1025. https://doi.org/10.1086/710625

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

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

Rothermund, K., & Koole, S. L. (2020). Rethinking emotion science: New theory section for Cognition & Emotion. *Cognition and Emotion*, *34*, 628–632. https://doi.org/10.1080/02699931.2020.1775924

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*, 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

S, K. (2017). Editorial. *Journal of Personality and Social Psychology*, *112*, 357–360. https://doi.org/10.1037/pspa0000077

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E. J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*, 211997. https://doi.org/10.1098/rsos.211997

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*, 107–114. https://doi.org/10.1177/2515245919838781

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*, 2295. https://doi.org/10.1002/icd.2295

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*, 25152459211007467. https://doi.org/10.1177/25152459211007467

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*, 744–755. https://doi.org/10.1177/1745691620966795

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

Schimmack, U. (2020, February 15). *Estimating the replicability of results in 'Journal of Experimental Social Psychology'*. https://replicationindex.com/2020/02/15/estrep-

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2019.1645

Schmidt, G. B., & Landers, R. N. (2013). Solving the replication problem in psychology requires much more than a website. *Industrial and Organizational Psychology*, *6*(3), 305–309. https://doi.org/10.1111/iops.12056

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*, 90–100. https://doi.org/10.1037/a0015108

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2015). Landy and Goodwin (2015) confirmed most of our findings then drew the wrong conclusions. *Perspectives on Psychological Science*, *10*, 537–538. https://doi.org/10.1177/1745691615589078

Schneider, D. J. (1992). Publication games: Reflections on Reis and Stiller. *Personality and Social Psychology Bulletin*, *18*, 498–503. https://doi.org/10.1177/0146167292184016

Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2020). Do open-science badges increase trust in scientists among undergraduates, scientists, and the public? *Psychological Science*, *33*, 1588–1604. https://doi.org/10.1177/09567976221097499

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515–530. https://doi.org/10.1037/0022-3514.51.3.515

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. https://doi.org/10.1037//0033-2909.105.2.309

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *BMJ (Clinical Research Ed*, *325*(7375), 1304–1304. https://doi.org/10.1136/bmj.325.7375.1304

Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F., & Wanyenze, R. K. (2018). Open data sharing and the Global South—Who benefits? *Science*, *359*, 642–643. https://doi.org/10.1126/science.aap8395

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572–582. https://doi.org/10.1037/a0034177

Sherman, J. W., & Rivers, A. M. (2021). There's nothing social about social priming: Derailing the 'train wreck'. *Psychological Inquiry*, *32*, 1–11. https://doi.org/10.1080/1047840X.2021.1889312

Sherman, R. C., Buddie, A. M., Dragan, K. L., End, C. M., & Finney, L. J. (1999). Twenty years of PSPB: Trends in content, design, and analysis. *Personality and Social Psychology Bulletin*, *25*, 177–187. https://doi.org/10.1177/0146167299025002004

Silan, M., Adetula, A., Basnight-Brown, D. M., Forscher, P. S., Dutra, N., & IJzerman, H. (2021). Psychological science needs the entire globe, part 2. *APS Observer*, *34*.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., & 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

Silverman, I. (1971). Crisis in social psychology: The relevance of relevance. *American Psychologist*, *26*, 583. https://doi.org/10.1037/h0031445

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*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SPSP Dialogue*, *4*. https://doi.org/10.2139/ssrn.2160588

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*, 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

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

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*, 87–104. https://doi.org/10.1177/1088868306294789

Smith, P. B., & Bond, M. H. (2022). Four decades of challenges by culture to mainstream psychology: Finding ways forward. *Journal of Cross-Cultural Psychology*, *53*(7–8), 729–751. https://doi.org/10.1177/00220221221084041

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*, 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Srivastava, S. (2012, September 27). *A pottery barn rule for scientific journals*. https://thehardestscience.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2022). Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods*, *13*, 88–108. https://doi.org/10.1002/jrsm.1529

Stanley, T. D., Doucouliagos, H., Ioannidis, J. P., & Carter, E. C. (2021). Detecting publication selection bias through excess statistical significance. *Research Synthesis Methods*, *12*, 776–795. https://doi.org/10.1002/jrsm.1512

Sternberg, R. S., & Sternberg, K. (2010). *The psychologist's companion: A guide to scientific writing for students and researchers* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511762024

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*, 1119–1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*, 479–491. https://doi.org/10.1017/S1930297500005611

Strack, F., & Schwarz, N. (2016). Editorial overview: Social priming: Information accessibility and its consequences. *Current Opinion in Psychology*, *12*. https://doi.org/10.1016/j.copsyc.2016.11.001

Stroebe, W. (2019). What can we learn from Many Labs replications. *Basic and Applied Social Psychology*, *41*, 91–103. https://doi.org/10.1080/01973533.2019.1577736

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71. https://doi.org/10.1177/1745691613514450

Sue, S. (1999). Science, ethnicity, and bias: Where have we gone wrong? *American Psychologist*, *54*, 1070–1077. https://doi.org/10.1037/0003-066X.54.12.1070

Sullivan, D. (2020). Social psychological theory as history: Outlining the critical-historical approach to theory. *Personality and Social Psychology Review*, *24*, 78–99. https://doi.org/10.1177/1088868319883174

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*, 989–1004. https://doi.org/10.1037/a0019507

Syed, M. (2020). *Whither the "White control group "? On the benefits of a comparative ethnic minority psychology*. https://doi.org/10.31234/osf.io/n4p73

Syed, M., & Kathawalla, U. (2021). Cultural psychology, diversity, and representation in open science. In K. C. McLean (Ed.), *Cultural methodologies in psychology: Capturing and transforming cultures* (pp. 427–454). Oxford University Press. https://doi.org/10.1093/oso/9780190095949.003.0015

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, *16*, 717–724. https://doi.org/10.1177/1745691620966796

Szucs, D., & Ioannidis, J. P. (2017a). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), 2000797. https://doi.org/10.1371/journal.pbio.2000797

Szucs, D., & Ioannidis, J. P. (2017b). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*, 390. https://doi.org/10.3389/fnhum.2017.00390

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*, 1–11. https://doi.org/10.1038/s41597-021-00981-0

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*, *76*, 116–129. https://doi.org/10.1037/amp0000622

The Center for Scientific Integrity. (2018). In *ISSN* (pp. 2692–465). http://retractiondatabase.org/

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553–571. https://doi.org/10.1007/s10869-020-09695-3

Towse, J. N., Ellis, D. A., & Towse, A. S. (2021). Opening Pandora's box: Peeking inside psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, *53*, 1455–1468. https://doi.org/10.3758/s13428-020-01486-1

Tracy, J. L., Robins, R. W., & Sherman, J. W. (2009). The practice of psychological science: Searching for Cronbach's two streams in social-personality psychology. *Journal of Personality and Social Psychology*, *96*, 1206–1225. https://doi.org/10.1037/a0015173

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, *145*, 643–654. https://doi.org/10.1037/xge0000159

van Aert, R. C., Wicherts, J. M., & Assen, M. A. (2019). Publication bias examined in metaanalyses from psychology and medicine: A meta-meta-analysis. *PLOS ONE*, *14*, 0215052. https://doi.org/10.1371/journal.pone.0215052

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016a). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459.

https://doi.org/10.1073/pnas.1521897113

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016b). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences*, *113*, 4935–4936. https://doi.org/10.1073/pnas.1609700113

van den Akker, O. R., Assen, M. A. L. M., & Bakker, M. (2023). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*, 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

van den Akker, O. R., Wicherts, J. M., Alvarez, L. D., Bakker, M., & Assen, M. A. (2023). How do psychology researchers interpret the results of multiple replication studies? *Psychonomic Bulletin & Review*, *30*, 1609–1620. https://doi.org/10.3758/s13423-022-02235-5

van der Steen, J. T., Bogert, C. A., Soest-Poortvliet, M. C., Fazeli Farsani, S., Otten, R. H., Ter Riet, G., & Bouter, L. M. (2018). Determinants of selective reporting: A taxonomy based on content analysis of a random selection of the literature. *PLOS ONE*, *13*, 0188247. https://doi.org/10.1371/journal.pone.0188247

van Rooij, I. (2019). Psychological Science needs theory development before preregistration. *Psychonomic Society Featured Comment*.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology-A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

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*(3), 1–5. https://doi.org/10.1525/collabra.13

Vasilevsky, N. A., Hosseini, M., Teplitzky, S., Ilik, V., Mohammadi, E., Schneider, J., & Holmes, K. L. (2021). Is authorship sufficient for today's collaborative research? A call for contributor roles. *Accountability in Research*, *28*, 23–43. https://doi.org/10.1080/08989621.2020.1779591

Vitelli, R. (1988). The crisis issue assessed: An empirical analysis. *Basic and Applied Social Psychology*, *9*, 301–309. https://doi.org/10.1207/s15324834basp0904_5

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*, 1628–1639. https://doi.org/10.1037/xge0000663

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem. *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

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. In *Advances in Methods and Practices in Psychological Science* (pp. 166–173). https://doi.org/10.1177/2515245918757427

Wegener, D. T., Fabrigar, L. R., Pek, J., & Hoisington-Shaw, K. (2022). Evaluating research in personality and social psychology: Considerations of statistical power and concerns about false findings. *Personality and Social Psychology Bulletin*, *48*, 1105–1117. https://doi.org/10.1177/01461672211030811

Wegner, D. M., Wenzlaff, R., Kerker, R. M., & Beattie, A. E. (1981). Incrimination through innuendo: Can media questions become public answers? *Journal of Personality and Social Psychology*, *40*, 822–832. https://doi.org/10.1037/0022-3514.40.5.822

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*, 267–295. https://doi.org/10.1037/emo0000226

Weigold, A., & Weigold, I. K. (2022). Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, *40*, 1302–1322. https://doi.org/10.1177/08944393211006847

West, S. G., & Gunn, S. P. (1978). Some issues of ethics and social psychology. *American Psychologist*, *33*, 30–38. https://doi.org/10.1037/0003-066X.33.1.30

Wible, J. R. (1992). Fraud in science an economic approach. *Philosophy of the Social Sciences*, *22*, 5–27. https://doi.org/10.1177/004839319202200101

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*, 202–217. https://doi.org/10.1037/teo0000137

Wilkinson, L., & Statistical Inference, T. F. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213802. https://doi.org/10.1177/25152459231213802

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*, 454–463. https://doi.org/10.1177/1948550619877412

Woolley, K., & Fishbach, A. (2017). Immediate rewards predict adherence to long-term goals. *Personality and Social Psychology Bulletin*, *43*, 151–162. https://doi.org/10.1177/0146167216676480

Wu, J., O'Connor, C., & Smaldino, P. E. (in press). The cultural evolution of science. In J. Kendal, R. Kendal, & J. Tehrani (Eds.), *The Oxford Handbook of Cultural Evolution*. Oxford University Press. https://osf.io/2ekcr/download.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, *e1*, 1–78. https://doi.org/10.1017/S0140525X20001685

Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L., & Tahk, A. M. (2019). Moderation of classic social psychological effects by demographics in the U.S. adult population: New opportunities for theoretical advancement. *Journal of Personality and Social Psychology*, *117*, 84–99. https://doi.org/10.1037/pspa0000171

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, 1–13. https://doi.org/10.1017/S0140525X17001972

# ENDNOTES



1. One example being Krosnick et al., 1993, published in *Journal of Social and Personality Psychology* (*JPSP*), who tried to discern consistent patterns of intercorrelation among attitude strength measures but found none. ↑

2. The high-profile case of social psychologist Diederik Stapel's large-scale data fabrication also became public knowledge that year (Budd, 2013), and investigations unearthed a handful of other presumed frauds in social and consumer psychology around that time. But while the Stapel case added urgency to resolving the methodological issues of the time, it did not directly impact the issues we are discussing. These issues involve practices used by authors who thought they were reporting their results normally and justifiably. Other means beyond the scope of this chapter are needed to discourage, detect, and punish intentional fabrication. ↑

3. "Replication" is ambiguous in that it can refer either to the attempt at replication, or the successful replication, however defined. "Non-replicators" here means the people who try and do not get significant results, not the people who never try at all. ↑

4. Checking whether results are correct and whether they can be independently derived from the original paper's data set is properly termed "reproducibility" despite some ambiguous usage of this word in the 2010's to refer to replication. See Nosek et al., 2022. ↑

5. These suspicions of bias have, sadly, sometimes been confirmed through the working of moralized group dynamics in the reform debates. Namely, there are several examples where researchers' skepticism about reform has been taken as a reason to subject their research to scrutiny – lab audits, replication, plausibility analyses – bringing to mind the corrupt use of tax agency investigations against political opponents. ↑