PERSPECTIVE

Science is not a signal detection problem

Brent M. Wilson[a,1], Christine R. Harris[a], and John T. Wixted[a,1]

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved February 6, 2020 (received for review August 19, 2019)

The perceived replication crisis and the reforms designed to address it are grounded in the notion that science is a binary signal detection problem. However, contrary to null hypothesis significance testing (NHST) logic, the magnitude of the underlying effect size for a given experiment is best conceptualized as a random draw from a continuous distribution, not as a random draw from a dichotomous distribution (null vs. alternative). Moreover, because continuously distributed effects selected using a $P < 0.05$ filter must be inflated, the fact that they are smaller when replicated (reflecting regression to the mean) is no reason to sound the alarm. Considered from this perspective, recent replication efforts suggest that most published $P < 0.05$ scientific findings are "true" (i.e., in the correct direction), with observed effect sizes that are inflated to varying degrees. We propose that original science is a screening process, one that adopts NHST logic as a useful fiction for selecting true effects that are potentially large enough to be of interest to other scientists. Unlike original science, replication science seeks to precisely measure the underlying effect size associated with an experimental protocol via large-*N* direct replication, without regard for statistical significance. Registered reports are well suited to (often resource-intensive) direct replications, which should focus on influential findings and be published regardless of outcome. Conceptual replications play an important but separate role in validating theories. However, because they are part of NHST-based original science, conceptual replications cannot serve as the field's self-correction mechanism. Only direct replications can do that.

replication crisis | null hypothesis significance testing | signal detection theory

Scientists generally conduct experiments in an effort to separate ideas that are true from ideas that are false, but many have raised concerns about how much progress is being made on that front. For example, Ioannidis (1) famously claimed that most published research findings are false, and Simmons et al. (2) reinforced that message by showing how easy it is for false hypotheses to yield statistically significant results. When prominent published results from fields as wide ranging as preclinical cancer trials (3), gene association studies (4), and social psychology (5) failed to replicate, alarm bells began to go off.

The idea that nonreplicable findings might be pervasive gained traction when the Open Science Collaboration (OSC2015) (6) attempted to replicate 100 representative psychology experiments, 97 of which originally achieved $P < 0.05$. To the surprise of many, only 36% of those 97 original experiments achieved $P < 0.05$ on replication, and the replicated effect sizes were, on average, only half the size of the original effect sizes (OSC2015) (table 1 of ref. 6). That pattern of results cemented the idea of a "replication crisis," and it has shaken the public's faith in science.

Had the large majority of those studies replicated at $P < 0.05$, with average effect sizes similar to the originally reported effect sizes (with ∼50% larger than the original effect size and ∼50% smaller), most would probably agree that science is functioning as it should be. However, such an outcome would mean that something went seriously wrong with the replication effort. Unless the original studies had 100% power (an obviously unrealistic assumption), original findings selected using a $P < 0.05$ filter must, on average, be associated with inflated effect sizes (7). Therefore, on replication, regression to the mean must occur. The original $P < 0.05$ studies did not have to be replicated to know, with virtual certainty, that their effect sizes would decline (Fig. 1). Fortunately, they did decline.

[a]Department of Psychology, University of California San Diego, La Jolla, CA 92093

Author contributions: B.M.W. and J.T.W. performed research; and B.M.W., C.R.H., and J.T.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: b6wilson@ucsd.edu or jwixted@ucsd.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1914237117/-/DCSupplemental.

First published March 3, 2020.

How much would the effect sizes be expected to decline given that the original experiments were associated with less than 100% power and their reported effects were selected using a $P < 0.05$ filter? The answer to that question is unknown, and that is precisely the point. The perceived replication crisis consists of comparing the observed OSC2015 outcome with an unrealistic outcome. To assess the state of science in light of the OSC2015 replication findings, it is important to first estimate the expected outcome given inevitable regression to the mean. We offer one such estimate here. Before delving into the quantitative details of that story, we first preview our findings and present what we believe to be a novel and practical vision of science. Indeed, it was our inquiry into regression to the mean, not our estimate of its magnitude, that led us to the vision we outline next.

Conceptualizing Science in Light of OSC2015

Because the role played by regression to the mean is unknown, there are two competing interpretations of the OSC2015 findings. The first is that most reported $P < 0.05$ findings in psychology are false (the veritable definition of a replication crisis). The second is that, even if true, most $P < 0.05$ effects are likely to be smaller than the original results suggest. This interpretation is harder to reconcile with the notion of a full-scale replication crisis because it assumes that findings in the scientific literature are generally true. Interestingly, this second interpretation was advanced by the authors of OSC2015 themselves in their response to critics (8). As they put it, "[t]he combined results of OSC2015's five indicators of reproducibility suggest that, even if true, most effects are likely to be smaller than the original results suggest" (ref. 8, p. 1037-c).

So which is it, interpretation 1 (most original $P < 0.05$ findings are false) or interpretation 2 (the original $P < 0.05$ findings are true, but their effect sizes are smaller than originally reported)? Intuition will not take us very far because this question cannot be answered without first answering another question that is rarely considered: what is the distribution of underlying (i.e., population) effect sizes associated with experiments conducted by scientists? It is important to be crystal clear about what this distribution of ground truth represents.

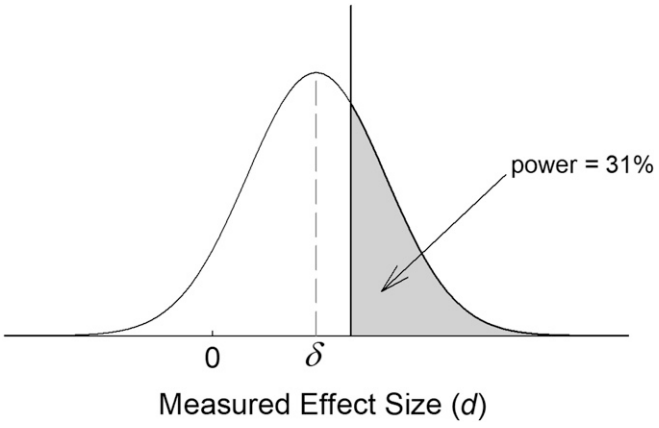

Fig. 1. Whenever power is less than 100%, the average of observed effect sizes (d) selected using a $P < 0.05$ filter would have to be greater than the underlying effect size, δ. For example, if power was as low as 31%, then all of the selected effect sizes (shaded region) would exceed the true underlying effect size of δ. Thus, when those effects were later replicated, regression to the mean would be observed.

Imagine the last 1,000 experiments conducted by 1,000 different scientists in a given field. Regardless of the observed effect sizes (e.g., Cohen's d) in these experiments and regardless of whether they achieved $P < 0.05$, they all had some true, underlying effect size (e.g., Cohen's δ). If we could somehow know what those 1,000 underlying effect sizes were, what would that distribution look like? We call this the prior distribution of underlying effect sizes, and the critical issue is whether that distribution is dichotomous, as assumed by null hypothesis significance testing (NHST) logic, or continuous (Fig. 2). Later in this article, we present a principled argument for assuming that the prior distribution is exponential in form. This distribution holds that underlying effect sizes range continuously across experiments from zero to ∞, with the direction of the underlying effect defined to be positive (Fig. 2*C*).

After proposing the exponential prior distribution of underlying effect sizes, we used it to simulate both the original and replication effect sizes reported in OSC2015. We ultimately conclude from these simulations and from additional analyses of the OSC2015 data that interpretation 2 is much more defensible than interpretation 1. In other words, our results are consistent with the idea that the original $P < 0.05$ findings in psychological science are generally "true" in the NHST sense of that word. In the NHST sense, a finding is true if its underlying effect size is greater than zero (ranging from negligible to large) and in the right direction. Our results further suggest that their underlying effect sizes are substantially smaller than originally reported, with much of the decline being attributable to regression to the mean.* This possibility seems important to carefully consider before reforming science in an effort to enhance replicability (11–15).

Is Regression to the Mean a Problem That Needs to Be Fixed?

If the OSC2015 replication data largely reflect regression to the mean, the implication would be that the original studies had considerably less than 100% power. If so, it is tempting to jump straight to the conclusion that the field should fix this problem by substantially increasing sample size (N), thereby increasing statistical power in future studies. Indeed, this recommendation may be the most widely agreed on reform by those working to improve scientific practices (13, 16–18).

Increasing N to boost statistical power would make sense if science actually were a binary signal detection problem, with underlying effect sizes being either literally zero (the null hypothesis) or some definite quantity, μ, greater than zero (the alternative hypothesis). Given that picture of underlying reality, increasing N would increase the "hit rate" (i.e., power) while leaving the "false alarm rate" (i.e., the alpha level) unchanged (13) (*SI Appendix*). However, if underlying reality is continuous, then, unless NHST were also abandoned, this cure would be worse than the disease. The reason is that conducting studies with ever-larger N will introduce ever-smaller underlying effect sizes into the $P < 0.05$ literature. Thus, in a manner of speaking, the false alarm rate (i.e., the rate of small underlying effects achieving statistical significance) will also increase as N increases.

*This issue has occasionally been pointed out, although without modeling the data (9, 10). Indeed, Trafimow (10) noted that "one way to view the Open Science Collaboration finding is that it provides empirical confirmation that psychology results are not immune to statistical regression" (p. 1190).

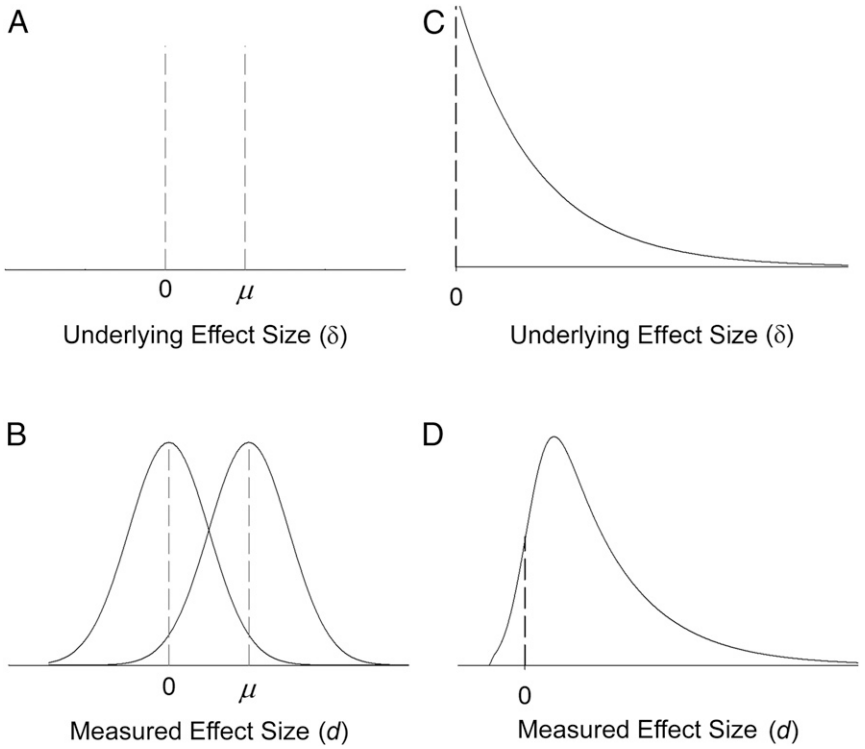

Fig. 2. (*A*) The prior distribution of underlying effect sizes (δ) assumed by standard NHST logic. (*B*) The distribution of measured effect sizes (*d*) after adding Gaussian measurement error. (*C*) The prior distribution of δ according to the continuous view using the exponential distribution as an example. (*D*) If the prior distribution of δ is exponential, the distribution of *d* (which includes Gaussian measurement error) would be an ex-Gaussian.

Reject the Null Hypothesis

As noted long ago by Meehl (19), with large-enough *N*, virtually every study would yield a significant result. In his words, "it is highly unlikely that any psychologically discriminable stimulation which we apply to an experimental subject would exert literally zero effect upon any aspect of his performance" (ref. 19, p. 109). Cohen (ref. 20, p. 1308) made a similar point

> The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection.

Echoing the same theme, Tukey (21) admonished that "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place" (ref. 21, p. 100).

The point is that the small effect that exists in every experiment (often for "nuisance" reasons having nothing to do with the theory the experimenter has in mind) would be detected with large-enough *N*. A troubling implication is that the detected effect would be in accordance with an a priori prediction made by a false theory half the time (19). Thus, maximizing *N* could be considered the ultimate questionable research practice, not the solution to what ails science. Rather than trying to fix the problem of regression to the mean by maximizing *N*, we submit that a much better solution would be for scientists and consumers of science alike to change their understanding of what a $P < 0.05$ finding actually means.

A $P < 0.05$ finding should not be regarded as a scientifically established discovery; instead, it should be regarded as a provisional finding, one that is likely in the right direction but with an observed effect size that is inflated to an unknown degree. Provisional findings merit the attention of scientists, but they have not yet been scientifically established for wider consumption. This is especially true if the finding is a surprising one. The original finding has to remain provisional until an independent laboratory directly replicates the experiment and obtains a precisely estimated effect size, one that is large enough to matter. Replication, after all, is the self-correction mechanism of science.

The Self-Correction Mechanism of Science

High-profile $P < 0.05$ findings are often surprising findings, and therefore, they understandably attract attention. The apparent demonstration of extrasensory perception (ESP) by Bem (22) is an extreme example. However, direct replications of those ESP experiments by independent laboratories failed to reproduce the originally observed effect (23, 24). Many interpret this episode as Engber (25) did in an article for *Slate* magazine entitled "Daryl Bem proved ESP is real. Which means science is broken," but we see the issue in a different light.

In our view, science is not broken because findings supporting ESP found their way into the literature only to be quickly corrected. This is precisely how science should work (i.e., science is supposed to correct its inevitable mistakes in due course). Instead, the problem is that, unlike the provisional findings reported by Bem (22), other high-profile findings sometimes appeared to pass the test of having been independently replicated, yet they ultimately turned out to have effect sizes very close to zero (26). Thus, something went wrong with the self-correction mechanism.

What went wrong, exactly? According to Pashler and Harris (5), the answer is that scientists assumed that theory-based conceptual replications were an adequate and perhaps even preferable substitute for direct replications. However, an often overlooked point is that effect sizes pertain to experimental protocols and have nothing to do with theory (i.e., theories do not have effect sizes). The self-correction mechanism consists of repeating the original experimental protocol as closely as possible, focusing on the method section of the original article, not on the theory that was tested (27, 28). The question of interest is whether an independent laboratory, duplicating the original experimental protocol as closely as possible but with much larger N, finds an effect size large enough to be worth factoring it into our understanding of the world (even if the effect is smaller than originally reported, as it is likely to be). As noted by Popper (ref. 29, p. 45),

> Indeed the scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed. No serious physicist would offer for publication, as a scientific discovery, any such 'occult effect,' as I propose to call it—one for whose reproduction he could give no instructions.

Whether a precisely measured replicated effect size is large enough to matter is not provided by the outcome of any statistical test but is instead a judgment call that will eventually be made by a consensus of scientists. As an example, with regard to the experimental protocols that initially yielded evidence for "money priming," a consensus seems to be emerging that any such effect is too close to zero to matter (30).

Conceptual replications, by contrast, pertain to theories, not to experimental protocols, and they serve a completely different function (28). This point is worth emphasizing because direct and conceptual replications are often pitted against each other as if the field should choose one or the other. However, both are important. Whereas direct replications indicate whether the original experimental protocol again yields a nonnegligible effect size, conceptual replications indicate whether the relevant theory again predicts the outcome when a different experimental protocol is used (31–33). The more often a theory correctly predicts the outcome of conceptual replications, the more faith one should have in that theory. However and this is the key point, conceptual replications do not constitute the self-correction mechanism of science. The reason is that a failed conceptual replication cannot be interpreted to mean that independent laboratories are unable to reproduce the originally reported finding (34). Instead, it might simply mean that the methodological departures from the original study pushed a valid theory beyond its domain of application. Thus, despite the somewhat misleading name, conceptual "replications" are part of original science (where provisional findings are published), not replication science.

NHST for Original Science vs. Replication Science

NHST has been excoriated for decades, with the main concern being its emphasis on either/or decision making (35–37). Here, we pile on by adding our own critique, making the argument that NHST is grounded in a false binary depiction of underlying reality. Even so, scientists should–as they long have–embrace it.

NHST Is a Useful Fiction for Original Science. If its foundational assumption is wrong, then what purpose does NHST serve? We suggest that its value can be appreciated by considering the underlying effect sizes (not the inflated observed effect sizes) that end up in the $P < 0.05$ literature. Specifically, as we detail later, applying the fiction of NHST to original science provides a mechanistic way to maximize the mean of the underlying effect sizes in the $P < 0.05$ literature, leaving behind an unholy mess of mostly small underlying effect sizes associated with hypotheses that few people care about except the scientists who dreamed them up in the first place. Original science relies on NHST precisely to serve that invaluable screening function. Note how different this goal is from the more intuitively appealing but less precise goal of trying to ensure that published $P < 0.05$ findings are true.

According to the case we lay out, the mean of the distribution of underlying effect sizes in the $P < 0.05$ literature is maximized (counterintuitively) by neither minimizing nor maximizing N; instead, it is maximized by optimizing N at an intermediate value. In other words, it is optimized by conducting original NHST scientific research in a way that is not radically different from how it is currently conducted. Using an intermediate value of N, original science screens for relatively large effect sizes on average.

There Is No Place for NHST in Replication Science. In contrast to the screening function served by original science, replication science is a truth-establishing endeavor. For experimental protocol i, the magnitude of δ_i is the only truth there is. Thus, replication science necessarily abandons the otherwise useful either/or fiction of NHST. The goal of replication science is to precisely measure the underlying effect size associated with a given experimental protocol (i.e., to quantify δ_i), and at this stage of the scientific process, maximizing N facilitates that goal (ref. 38 is a recent example).

Ideally, the replication would be a registered report reviewed in advance and published regardless of outcome (39, 40). Otherwise, "failures to replicate" might become easier to publish than successful replications or vice versa. However, in contrast to original science, publication bias for replication science is unjustifiable, which is why registered replications seem like the way to go.

Finally, a large-N direct replication is resource intensive, potentially involving many laboratories. Thus, it makes sense to reserve those resources to replicate original $P < 0.05$ findings that matter (i.e., published $P < 0.05$ findings that gain currency), not every $P < 0.05$ finding. Indeed, the more important the finding, the more sense it makes to devote extensive resources to the replication effort, perhaps going so far as to employ "radical randomization" (41). Given its focus on large-N replications of a relatively small proportion of published findings (namely, the important ones), our vision is almost the diametric opposite of the "large-N, publish everything" vision of science advocated by others (36).

The foundation of our vision of science comes from a consideration of a theoretically plausible distribution of underlying effect sizes associated with the experiments that scientists conduct, some of which end up in the $P < 0.05$ literature (with inflated observed effect sizes). We now turn to a more detailed inquiry into the underlying reality of science.

Science Is Not a Signal Detection Problem

NHST assumes that the underlying effect size (δ) is a discrete, binary variable to which measurement error is added (Fig. 2*B*). Anyone who has ever performed a power analysis before conducting an experiment has come into contact with this binary view of underlying effect sizes. The familiar steps are as follows: 1) specify the effect size associated with the null hypothesis (often $\delta = 0$), 2) specify the effect size associated with the alternative hypothesis (e.g., $\delta = 0.30$), 3) select an alpha level (usually 0.05), 4) select

desired power (e.g., 0.80), and 5) use a power calculator to determine the necessary N. In this approach, underlying effect sizes are assumed to correspond to a strictly dichotomous distribution as if δ is equal to either 0 or 0.30, but effect sizes of 0.20 or 0.40 are so unthinkable as to not even be worth mentioning, much less be taken seriously.

Contrary to what we pretend to be true when computing statistical power, effect sizes are better conceptualized as having been drawn from a nonbinary continuous distribution. As noted earlier, this perspective inherently rejects the idea that science is a signal detection problem. Ironically, signal detection theory evolved from the NHST approach of Fisher (42) ("false" corresponds to $\delta = 0$; true corresponds to the rejection of that idea) as elaborated by Neyman and Pearson (43), who proposed also taking into account the specific magnitude of the alternative hypothesis (44, 45). However, a key feature of an actual signal detection problem is that each trial can be unambiguously categorized as a stimulus-present trial or a stimulus-absent trial. It is not an assumption; it is literally true because an intelligent agent (namely, the experimenter) has arranged the task to be that way. In NHST, by contrast, the idea that δ is either true (i.e., $\delta \neq 0$) or false ($\delta = 0$) is simply a formalization, one adopted for its utility, not to accurately model underlying reality. In reality, δ is almost certainly a continuous variable, and no experiment needs to be performed to reject the null hypothesis of $\delta = 0$ because it can be safely rejected a priori (19–21).

What Is the Prior Distribution of δ?

Although we know almost nothing about the prior distribution of underlying effect sizes, we do know something about it. We know, for example, that δ ranges from zero to infinity, with the direction of the effect defined as positive. We also have some information about its mean. For example, when OSC2015 replicated 97 representative $P < 0.05$ effects from experimental psychology, the mean of the absolute values of the replication effect sizes—which provides a relatively unbiased estimate of the mean of their underlying effect sizes—was approximately $\bar{d} \approx 0.60$ (46).[†] Thus, it seems reasonable to suppose that the mean of the prior distribution of the underlying effect sizes associated with all psychology experiments, including the nonsignificant effects that were never published, would be less than that, perhaps ~0.30.

If all we know about a distribution is 1) its range and 2) its mean, then as noted by Jaynes (47), the maximum entropy distribution—that is, the distribution that is "maximally noncommittal with regard to missing information" (ref. 47, p. 623)—is the exponential. In truth, we do not know the exact mean of the underlying effect size distribution. However, given how much of the infinite range of possibilities we can safely rule out, it seems reasonable to proceed as though we do. We therefore used the exponential as the prior distribution of underlying effect sizes (as illustrated earlier in Fig. 2*C*). The underlying effect size for any given experiment, i, is conceptualized as a random draw from this distribution.

Simulating Scientific Research

To investigate the implications of the nonbinary view of science, we modeled the OSC2015 data—both the original experiments and the replication experiments—via simulation. At a minimum, our study provides an existence proof that regression to the mean can result in substantially reduced effect sizes on replication, just as Simmons et al. (2) provided an existence proof showing that flexibility in data analysis could increase scientific false positives. However, our simulation provides more than just an existence proof because it is, additionally, constrained by empirical data.

Original Experiments. As illustrated in Fig. 3, the simulation of original experiment, i, involved 1) a random draw, δ_i, from an underlying exponential effect size distribution with mean $\bar{\delta}$ and 2) a random draw from a conceptually related sample size distribution governed by a parameter g, yielding sample size, N_i (details are in *SI Appendix*). Next, for each simulated subject, j, random error drawn from a unit normal distribution was independently added to δ_i to create individual x_{ij} scores. A one-sample t test was then performed on those data, and the observed Cohen's d effect size was computed from that value using the formula $d_i = \frac{t_i}{\sqrt{N_i}}$. This process was repeated for a large number of simulated experiments.

The two free parameters ($\bar{\delta}$ and g) (Fig. 3) were adjusted separately for the cognitive and social psychology experiments until the simulated $P < 0.05$ data approximately matched 1) the P curves (48) for the original experiments replicated by OSC2015 and 2) the mean of the observed Cohen's d effect size distributions for the $P < 0.05$ original experiments replicated by OSC2015. With regard to the estimated mean of the prior distribution of underlying effect sizes, the final parameter estimates for cognitive and social psychology were $\bar{\delta}_{Cog} = 0.53$ and $\bar{\delta}_{Soc} = 0.22$, respectively. In other words, the data suggest that the underlying effect sizes in cognitive psychology are larger than those in social psychology.

Replication Experiments. For the subset of simulated original experiments yielding statistically significant observed effect sizes ($P < 0.05$, two tailed), we performed simulated replication experiments, generating another set of observed effect sizes. Each simulated replication experiment was based on the same underlying effect size (δ_i) used for the corresponding original experiment, but the sample size was determined using a power calculator. More specifically, based on the observed effect size (d_i) of the simulated original study, N_i for the simulated replication experiment was selected to achieve 90% power (following the practice used for real data in OSC2015). In the end, we had one set of d_i values from the simulated original studies and a corresponding set of d_i values from the simulated replication experiments.

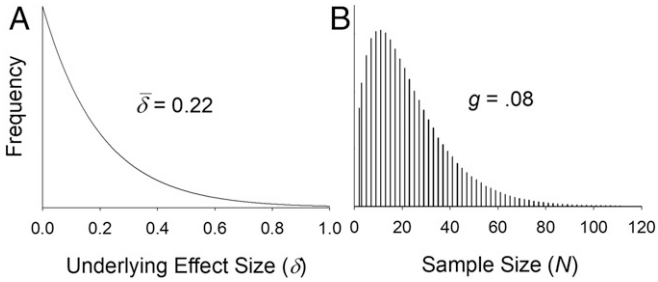

Fig. 3. Example of the underlying exponential effect size distribution (*A*) and sample size distribution (*B*) used for our simulated experiments. These examples were taken from our simulation of social psychology, where the mean of the effect size distribution is 0.22 and the mean of the sample size distribution (with g set to 0.08) is ~23.

[†]OSC2015 replicated 97 $P < 0.05$ effects as reported in table 1 of ref. 6, but the original and replicated effect sizes are available for only 94 of them. Thus, from here on, we consider those 94 effects.

Estimated Regression to the Mean. Our initial simulation based on the model specified above indicated that ~70% of the reduction in the observed effect sizes in OSC2015 may reflect regression to the mean. Model variations involving plausible (albeit subjective) assumptions yielded lower estimates falling in the 40 to 50% range (*SI Appendix*). Although the exact estimate will vary depending on the model that one finds to be the most compelling, our results lend credibility to the idea that regression to the mean played a substantial role in OSC2015.

Regression to the mean would potentially account for all of the observed effect size decline if the original studies involved no questionable research practices (QRPs) and the replication studies were all perfect replicas of the original studies. However, it seems unlikely that QRPs were completely absent in the original studies and that the replication studies were all perfect replicas of the original studies. After all, when asked, scientists admit to sometimes engaging in QRPs (49). Doing so must have some effect on the observed outcome. Similarly, even a photocopy of a document is not a perfect replica of the original because some degree of replication error is inevitable. In the OSC2015 replications, for example, the procedure was modified—in several cases dramatically so according to some—in order to perform the study in countries with cultures and languages different from the original studies (50).

More generally, additional factors, such as flexibility in experimental design and data analysis, can also lead to a reduction in replication effect sizes over and above regression to the mean (51). Nevertheless, beyond a variety of factors that may have played some role, our model-based simulation suggests that the OSC2015 replication results were substantially influenced by regression to the mean.

Note that, according to our simulation, both the original and replication effect sizes were larger for the experiments from cognitive psychology than social psychology. Also, simulated cognitive experiments were more likely to replicate at $P < 0.05$ (71%) than simulated social experiments (54%). However, despite superficial appearances, it does not automatically follow that cognitive psychology is a stronger science than social psychology (46, 52) (*SI Appendix*).

Most of the Original $P < 0.05$ Findings Are True, Not False

Overall, our analysis stands against a common intuitive understanding of the OSC2015 findings, which is that the 36% of experiments that successfully replicated at $P < 0.05$ are true positives ($\delta > 0$) and that the 64% of experiments that failed to replicate at $P < 0.05$ are false positives ($\delta = 0$). If one adopts the NHST view of what true means, then our model-based inquiry suggests that the OSC2015 findings are best characterized by interpretation 2 presented earlier: the original findings are generally true, but the effect sizes are smaller than originally reported.

Another way to make this point without relying on our model is to directly test whether the 64% of replication experiments with nonsignificant outcomes were (originally) false positives. Clearly, they were not because the distribution of the Cohen's *d* effect sizes from the nonsignificant replication studies was not centered on 0 but was instead significantly greater than 0: $t(59) = 3.47$, $P = 0.001$. Note that these findings were selected using a $P > 0.05$ filter and are therefore biased towards 0 (i.e., if they were replicated again, the now-unbiased effect-size estimates would be larger). Yet, even without correcting for that bias, the distribution of nonsignificant effect sizes from OSC2015 is significantly greater than 0. Thus, like the $P < 0.05$ effect sizes, many (if not all) of the nonsignificant effect sizes are also true in the NHST sense. Moreover, they would have been detected at $P < 0.05$ had the replication studies tested a much larger number of subjects (53). That fact underscores our point that increasing *N* to increase power will fill the scientific literature with ever smaller underlying effects.

In Defense of NHST

Against a relentless torrent of criticism that now spans more than half a century, only a few have come to the defense of NHST (54). Even so, NHST has remained the dominant approach to conducting research in many scientific fields. Why? We contend that, despite its flaws, NHST serves a useful purpose that none of the proposed alternative approaches has yet demonstrated.

What NHST Selects vs. What It Leaves Behind. Fig. 4*A* shows the exponential prior distribution of underlying effect sizes for the social psychology simulations described earlier ($\bar{\delta} = 0.22$). Fig. 4 *B* and *C* shows the distributions of underlying effect sizes for the simulated nonsignificant findings and significant findings, respectively. Finally, Fig. 4*D* shows the (inflated) observed effect sizes associated with the significant findings. Note that the observed effect sizes in Fig. 4*D* are always positive on the assumption that an experimenter who publishes a $P < 0.05$ finding in the wrong direction would be unaware of the sign error. The negative underlying effect sizes in Fig. 4 *B* and *C* (which were positive in Fig. 4*A*) indicate that the observed effect size in our simulation was in the wrong direction.

A large percentage of original experiments (78%) yields results that are not statistically significant (Fig. 4*B*). These filtered-out experiments had an average underlying effect size even closer to zero than the average underlying effect size of the prior distribution, and a high percentage of them has observed effect sizes that are in the wrong direction. Because they did not yield significant results, these studies would mostly end up in the "file drawer." By contrast, a smaller percentage of experiments (22%) yields a statistically significant outcome (Fig. 4*C*).[‡] These experiments have a considerably larger average underlying effect size relative to the prior distribution, and less than 5% are sign errors (55).

Fig. 4*C* illustrates the invaluable screening function served by NHST and publication bias. From a prior distribution with a mean underlying effect size of 0.22, a new distribution with a substantially higher mean (0.43) is selected for consideration by other scientists. However, these $P < 0.05$ findings reflect a distribution of underlying effect sizes, some of which are too close to 0 to matter and a few of which are slightly in the wrong direction. Because we do not know which observed effects fall into that near-0 region, $P < 0.05$ findings are not yet scientifically established discoveries. They are, however, worth considering, and those that happen to gain currency are worth directly replicating in a large-*N* investigation.

Publication bias is almost always construed as a cost, not a benefit, because nonsignificant findings end up in the file drawer. However, it seems important to distinguish between two separate drawers: 1) the file drawer, consisting of invisible direct replications of findings in the $P < 0.05$ literature that failed to yield a significant result and 2) the junk drawer, consisting of idiosyncratic, once-tested ideas that a researcher dreamed up and tested but that failed to yield a significant result. Everyone would like to see the file drawer become visible. However, if everything were

[‡]Power was, therefore, 22%, which might sound surprisingly low. However, translated into binary logic, imagine that the prior odds of an effect being true are 1:4, power is 80%, and the alpha level is 5%. In that seemingly reasonable scenario, only 20% of experiments would similarly yield a $P < 0.05$ result.

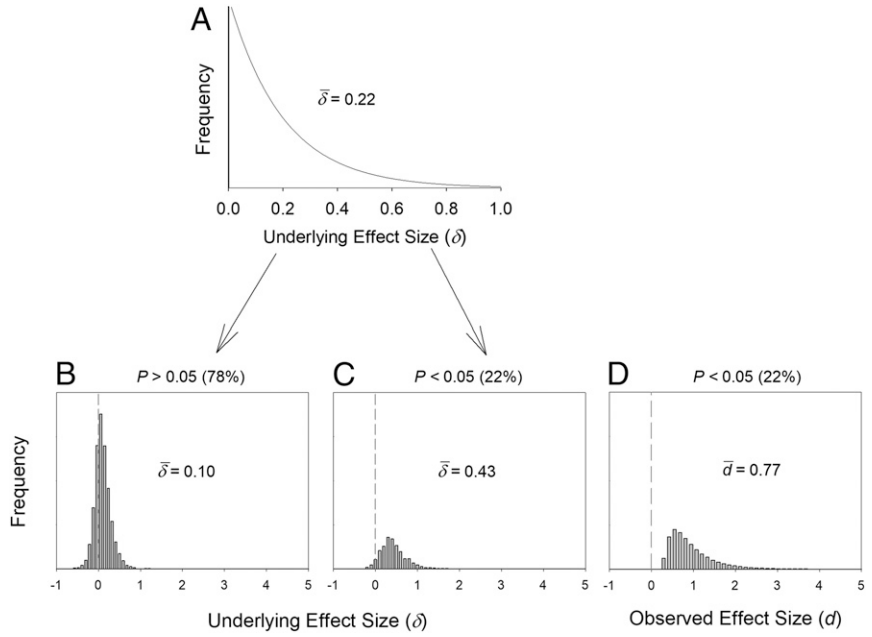

Fig. 4. (*A*) Prior distribution of underlying effect sizes used earlier for social psychology simulation. (*B*) Underlying effect sizes associated with nonsignificant outcomes in our simulation study (78% of the simulated experiments). (*C*) Underlying effect sizes associated with significant outcomes in our simulation study (22% of the simulated experiments). Of the significant effects, ∼3.5% are in the wrong direction relative to the observed effect size (shown here as negative underlying effects). (*D*) Observed effect sizes associated with significant outcomes in our simulation study.

published to accomplish that worthy goal, it would come at the cost of polluting the literature with the junk drawer as well (illustrated in Fig. 4*B*).

What, Exactly, Is the Goal of NHST Research? In a nonbinary effect size world, the goal of NHST research has to be specified with respect to a distribution of underlying effect sizes. In the example above, the mean of the distribution of underlying effect sizes that appear in the literature (Fig. 4*C*) is approximately double the mean of the prior distribution of underlying effect sizes (Fig. 4*A*). This outcome occurred using an average simulated *N* of ∼23. By contrast, if *N* were set to an extremely large value, then virtually every experiment would achieve $P < 0.05$, and the distribution of statistically significant underlying effect sizes would essentially reproduce the prior distribution with a mean of 0.22. That outcome (i.e., smaller underlying effect sizes in the statistically significant literature) illustrates a cost of maximizing *N* even if doing so did not consume additional resources, which it also does.

Interestingly, a similar result is obtained if *N* is minimized. That is, small-*N* studies are also counterproductive because they too minimize underlying effect sizes that appear in the $P < 0.05$ literature. When NHST is applied to a world in which the prior distribution is a continuous exponential, the mean of the underlying distribution of $P < 0.05$ effect sizes is maximized using an intermediate value of *N* (*SI Appendix*). Thus, a rational goal for original science is to use an intermediate sample size that maximizes the average δ associated with the published $P < 0.05$ findings (Fig. 5).[§]

[§]A reasonable alternative goal might be to choose the value of *N* that would maximize the mean of the underlying effect sizes subject to the constraint that the expected observed effect size is no more than, say, 1.5 times the expected underlying effect size.

An NHST Analogy. Imagine testing 100 candidates for a new basketball team. In this analogy, each candidate represents an experiment. The underlying ability to play basketball is a continuously distributed variable (it might even be exponential), and each candidate's ability represents an effect size. At considerable expense, we could test everyone's basketball-playing ability precisely, but we instead decide to use a much less costly screening test to separate the "true" basketball players from the "false" basketball players (a binary fiction we set up for the sake of convenience). Our screening test consists of selecting players who hit 10 free throws in a row, an outcome that represents achieving $P < 0.05$. Imagine that 5 of the 100 candidates pass our free-throw

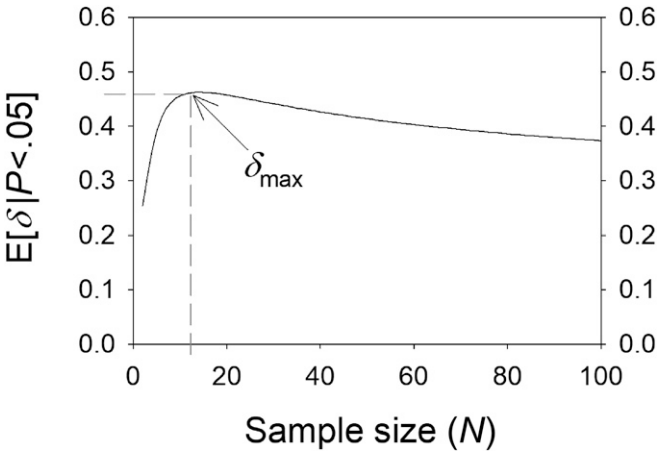

Fig. 5. Expected value of δ given a $P < 0.05$ outcome assuming an exponential prior with $\delta = 0.22$. The mean of the distribution of $P < 0.05$ underlying effect sizes is maximized at $\delta_{max} = 0.463$ when an intermediate value of *N* is used ($N = 13$ in this example). The expected value of the inflated observed effect size associated with δ_{max} (not shown in the figure) is $d = 0.850$.

test, so they make the team, an outcome that represents empirical findings being published.

The first point to make about these 5 players is that, on average, they are likely to be better than the 100 candidates we started with, a fact that highlights the value of our relatively inexpensive screening test. The second point to make is that if we tested these 5 players again they would likely hit fewer than 10 free throws in a row, on average (i.e., on replication, regression to the mean would almost certainly be observed). Thus, although these 5 players are good, on average, they are not as good as their performance on the original test made them seem. Still, if no one ever came to watch this new team play, the fact that some of them might not be very good would not matter very much. By contrast, if their perfect free-throw performance got the attention of recruiters for the national Olympic basketball team, it would now make sense to carefully assess each of the 5 selected players on their basketball playing ability even though the cost of doing so would be high. This analogy represents how NHST operates in a continuous effect-size world, and the key point is that it has nothing to do with true vs. false (except as part of a screening fiction). The specific magnitude of each player's basketball ability is the only truth there is.

Conclusion

The first sentence of Cohen's (56) treatise on NHST research reads as follows: "After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists" (ref. 56, p. 997). That paper was published 25 y ago, which means that NHST research has now survived more than six decades of severe criticism. Even the most ardent opponent of NHST research might agree that it is not completely irrational to assume that the next 65 y of scientific research are going to look a lot like the last 65 y of scientific research, namely lots of NHST research along with relentless criticism of it.

If we are right to assume that underlying effect sizes are continuously distributed and that NHST research will likely be with us for the foreseeable future, then what would the goal of scientific research be? The vision of science we set forth above holds that original science based on NHST should be viewed as a screening process aimed at other scientists, whereas replication science involving large-*N* direct replications without regard for statistical significance should be viewed as a confirmation process aimed at everyone (scientists, textbook writers, the media, policy makers, etc.).

An advantage of the screening-plus-confirmation vision of science is that the changes to current practices that would be required are relatively modest. That makes it a more feasible vision than alternative ideas that depend on much more sweeping changes (and on the untested hope that the benefits of those changes to science will outweigh the unintended consequences). In addition, it would be cost effective in that the resources required to conduct large-*N* direct replications would not be expended on every finding or even on every $P < 0.05$ finding. Instead, unlike the "large-*N*, publish everything vision" (36), resources would be concentrated on published findings that gain currency (57). This vision may not appeal to the metaanalyst, but in a resource-limited world, large-*N* direct replications of influential findings may provide the best path to the truth (58).

A headwind for our vision is that it depends on replication science actually happening. It therefore seems important for funding agencies to directly incentivize such work by funding proposals to replicate influential findings in the published literature. Given that NHST research can only be reasonably viewed as a screening process, not as a truth-establishing endeavor, major funding agencies should set aside a significant fraction of their budgets (e.g., 10% or more) for independent, large-*N* direct replications of influential findings.

In addition, outlets for replication science are essential, but not every journal is open to the idea. One approach might be for an original science journal to publish peer reviewed and fully indexed direct replications in an online sister journal (e.g., for the family of *Nature* journals, this new member of the family might be called *Nature Replications*). The online versions of the originally published studies would then be amended by adding conspicuous links to any and all large-*N* direct replications of it that are published in the sister journal.

In summary, for original science, NHST logic is a useful fiction, one that serves a screening function, much like medical screening does. By contrast, at the replication stage, we add our voice to the many who have called for the abandonment of NHST logic (36, 37). At that stage, the only relevant truth is the singular underlying effect size associated with the original experimental protocol. There is no other truth, and the goal of replication science should be to bring that truth to light as precisely as possible.

Data Availability. The code used for simulating the OSC2015 data is available at https://osf.io/pvxzs/.

Acknowledgments

We thank two anonymous reviewers for their careful evaluation and critique of our Perspective.

1 J. P. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
2 J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
3 C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
4 C. F. Chabris *et al.*, Most reported genetic associations with general intelligence are probably false positives. *Psychol. Sci.* **23**, 1314–1323 (2012).
5 H. Pashler, C. R. Harris, Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* **7**, 531–536 (2012).
6 Open Science Collaboration, PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
7 F. L. Schmidt, What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am. Psychol.* **47**, 1173–1181 (1992).
8 C. J. Anderson *et al.*, Response to comment on "Estimating the reproducibility of psychological science." *Science* **351**, 1037 (2016).
9 K. Fiedler, J. Prager, The regression trap and other pitfalls of replication science—illustrated by the report of the Open Science Collaboration. *Basic Appl. Soc. Psych.* **40**, 115–124 (2018).
10 D. Trafimow, An a priori solution to the replication crisis. *Philos. Psychol.* **31**, 1188–1214 (2018).
11 D. J. Benjamin *et al.*, Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
12 D. Bishop, Rein in the four horsemen of irreproducibility. *Nature* **568**, 435 (2019).
13 K. S. Button *et al.*, Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
14 B. A. Nosek, C. R. Ebersole, A. C. DeHaven, D. T. Mellor, The preregistration revolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2600–2606 (2018).

15 B. O. Turner, E. J. Paul, M. B. Miller, A. K. Barbey, Small sample sizes reduce the replicability of task-based fMRI studies. *Commun Biol* **1**, 62 (2018).
16 M. R. Munafo *et al.*, A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
17 J. B. Asendorpt *et al.*, Recommendations for increasing replicability in psychology. *Eur. J. Pers.* **27**, 108–119 (2013).
18 D. C. Funder *et al.*, Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Pers. Soc. Psychol. Rev.* **18**, 3–12 (2014).
19 P. E. Meehl, Theory testing in psychology and physics: A methodological paradox. *Philos. Sci.* **34**, 103–115 (1967).
20 J. Cohen, Things I have learned (thus far). *Am. Psychol.* **45**, 1304–1312 (1990).
21 J. W. Tukey, The philosophy of multiple comparisons. *Stat. Sci.* **6**, 100–116 (1991).
22 D. J. Bem, Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* **100**, 407–425 (2011).
23 J. Galak, R. A. LeBoeuf, L. D. Nelson, J. P. Simmons, Correcting the past: Failures to replicate ψ. *J. Pers. Soc. Psychol.* **103**, 933–948 (2012).
24 S. J. Ritchie, R. Wiseman, C. C. French, Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS One* **7**, e33423 (2012).
25 D. Engber, Daryl Bem proved ESP is real. Which means science is broken. *Slate*, 17 May 2017. https://web.archive.org/web/20191220053827/https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html. Accessed 20 December 2019.
26 D. Rohrer, H. Pashler, C. R. Harris, Do subtle reminders of money change people's political views? *J. Exp. Psychol. Gen.* **144**, e73–e85 (2015).
27 D. J. Simons, The value of direct replication. *Perspect. Psychol. Sci.* **9**, 76–80 (2014).
28 R. A. Zwaan, A. Etz, R. E. Lucas, M. B. Donnellan, Improving social and behavioral science by making replication mainstream: A response to commentaries. *Behav. Brain Sci.* **41**, e157 (2018).
29 K. R. Popper, *The Logic of Scientific Discovery* (Basic Books, Inc., 1959).
30 T. Chivers, What's next for psychology's embattled field of social priming. *Nature* **576**, 200–202 (2019).
31 C. S. Crandall, J. W. Sherman, On the scientific superiority of conceptual replications for scientific progress. *J. Exp. Soc. Psychol.* **66**, 93–99 (2016).
32 N. Schwarz, F. Strack, Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Soc. Psychol.* **45**, 305–306 (2014).
33 W. Stroebe, F. Strack, The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* **9**, 59–71 (2014).
34 S. Schmidt, Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* **13**, 90–100 (2009).
35 V. Amrhein, S. Greenland, B. McShane, Scientists rise up against statistical significance. *Nature* **567**, 305–307 (2019).
36 G. Cumming, The new statistics: Why and how. *Psychol. Sci.* **25**, 7–29 (2014).
37 B. B. McShane, D. Gal, A. Gelman, C. Robert, J. L. Tackett, Abandon statistical significance. *Am. Stat.* **73**, 235–245 (2019).
38 D. P. Morgan, J. Tamminen, T. M. Seale-Carlisle, L. Mickes, The impact of sleep on eyewitness identifications. *R. Soc. Open Sci.* **6**, 170501 (2019).
39 C. Chambers, What's next for registered reports? *Nature* **573**, 187–189 (2019).
40 D. S. Lindsay, Preregistered direct replications in psychological science. *Psychol. Sci.* **28**, 1191–1192 (2017).
41 B. Baribault *et al.*, Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2607–2612 (2018).
42 R. A. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, Scotland, 1925).
43 J. Neyman, E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. A* **231**, 289–337 (1933).
44 D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, NY, 1966).
45 J. T. Wixted, The forgotten history of signal detection theory. *J. Exp. Psychol. Learn. Mem. Cogn.* **46**, 201–233 (2020).
46 B. M. Wilson, J. T. Wixted, The prior odds of testing a true effect in cognitive and social psychology. *Adv. Methods Pract. Psychol. Sci.* **1**, 186–197 (2018).
47 E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
48 U. Simonsohn, L. D. Nelson, J. P. Simmons, p-Curve and effect size: Correcting for publication bias using only significant results. *Perspect. Psychol. Sci.* **9**, 666–681 (2014).
49 L. K. John, G. Loewenstein, D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532 (2012).
50 D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, Comment on "Estimating the reproducibility of psychological science." *Science* **351**, 1037 (2016).
51 C. J. Bryan, D. S. Yeager, J. M. O'Brien, Replicator degrees of freedom allow publication of misleading failures to replicate. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 25535–25545 (2019).
52 National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (The National Academies Press, Washington, DC, 2019).
53 S. E. Maxwell, M. Y. Lau, G. S. Howard, Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* **70**, 487–498 (2015).
54 J. Krueger, Null hypothesis significance testing. On the survival of a flawed method. *Am. Psychol.* **56**, 16–26 (2001).
55 A. Gelman, J. Carlin, Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).
56 J. Cohen, The earth is round ($p < 0.05$). *Am. Psychol.* **49**, 997–1003 (1994).
57 S. Lewandowsky, K. Oberauer, Low replicability can support robust and efficient science. *Nat. Commun.* **11**, 358 (2020).
58 A. Kvarven, E. Strømland, M. Johannesson, Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nat. Hum. Behav.*, 10.1038/s41562-019-0787-z (2019).