Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-022-02098-w

THEORETICAL/REVIEW

Theoretical false positive psychology

Brent M. Wilson[1] · **Christine R. Harris**[1] · **John T. Wixted**[1]

Accepted: 4 April 2022

Abstract
A fundamental goal of scientific research is to generate true positives (i.e., authentic discoveries). Statistically, a true positive is a significant finding for which the underlying effect size (δ) is greater than 0, whereas a false positive is a significant finding for which δ equals 0. However, the null hypothesis of no difference ($\delta = 0$) may never be strictly true because innumerable nuisance factors can introduce small effects for theoretically uninteresting reasons. If δ never equals zero, then with sufficient power, every experiment would yield a significant result. Yet running studies with higher power by increasing sample size (N) is one of the most widely agreed upon reforms to increase replicability. Moreover, and perhaps not surprisingly, the idea that psychology should attach greater value to small effect sizes is gaining currency. Increasing N without limit makes sense for purely measurement-focused research, where the magnitude of δ itself is of interest, but it makes less sense for theory-focused research, where the truth status of the theory under investigation is of interest. Increasing power to enhance replicability will increase true positives at the level of the effect size (statistical true positives) while increasing false positives at the level of theory (theoretical false positives). With too much power, the cumulative foundation of psychological science would consist largely of nuisance effects masquerading as theoretically important discoveries. Positive predictive value at the level of theory is maximized by using an *optimal N*, one that is neither too small nor too large.

Keywords Null hypothesis significance testing · False positives · Positive predictive value · Replication crisis

Introduction

Metascience has been defined as "turning the lens of science on itself" (Schooler, 2014). To enhance replicability, scientists have proposed a variety of methodological and statistical reforms, such as preregistering experiments (Nosek et al., 2018), conducting experiments with higher power (Button et al., 2013), and replacing null-hypothesis statistical testing with a measurement approach (Cumming, 2014). To further enhance the rigor of scientific research, reforms have also been proposed at the level of theory, such as making use of formal models to precisely specify assumptions about underlying theoretical processes (Borsboom et al., 2021; Muthukrishna & Henrich, 2019; Navarro, 2021; Oberauer & Lewandowsky, 2019). An argument could be made – and we do so here – that formal models can be used to enhance the rigor of metascience in much the same way.

Models make assumptions about latent variables, which are variables that cannot be directly observed (e.g., the strength of sensations, memories, emotions, etc.) or that are observable in principle but would be impractical to measure. An example of the latter is the underlying (i.e., true) effect size associated with an experimental protocol. In principle, the underlying effect size could be exactly measured by testing the entire population. However, because it would be impractical to do so, the underlying effect size for a given experimental protocol, like the strength of the sensation generated by a tone, is a latent variable.

If the magnitude of an underlying effect size for even one experimental protocol is essentially unknowable, trying to fathom how underlying effect sizes are distributed across the entire population of experiments that define a field might seem like a fruitless endeavor. However, the strength of a sensation generated by even one test stimulus is similarly unknowable, yet the distribution of the strength of sensations generated by test stimuli across trials has been profitably conceptualized in

✉ Brent M. Wilson
b6wilson@ucsd.edu

✉ John T. Wixted
jwixted@ucsd.edu

[1] Department of Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Published online: 02 May 2022

terms of signal detection theory since the dawn of experimental psychology (Fechner, 1860; Wixted, 2020). In the signal detection framework, the strength of a sensation is typically conceptualized as a random draw from a Gaussian distribution of sensations across trials. The magnitude of an underlying effect size for a to-be-conducted experiment can be similarly conceptualized as a random draw from an unknown parent distribution. It is worth trying to specify what that distribution might be because, once defined, it becomes possible to model the distribution of underlying effect sizes that end up in the scientific literature.

Our focus here is not on traditional statistical distributions, such as the *t*-distribution or its close relative, the distribution of Cohen's *d*. Distributions like these are often considered when analyzing data from a single experiment. Our focus is on the distribution of the true underlying effect sizes – such as the distribution of Cohen's delta (δ) – across a population of experiments, without measurement error. By contrast, the distributions of *t* or *d* represent statistics that would be observed if a single experiment (with a single δ) were repeated many times, and they reflect measurement error.

Just as different people have different heights, different experimental protocols have different underlying effect sizes. And just as a given person can be conceptualized as a random draw from the distribution of heights associated with a population of people, a given experiment can be conceptualized as a random draw from the distribution of underlying effect sizes associated with a population of experiments (e.g., from experimental psychology). What is the shape of that distribution?

We suppose that most researchers have not considered that question very deeply beyond the slight consideration given to it when they consider statistical power for a single experiment. To perform a power analysis, researchers who rely on traditional null-hypothesis significance testing (NHST) assume that the underlying effect size associated with the to-be-conducted experiment is either $\delta = 0$ (the null hypothesis of no difference) or some specific value greater than 0, such as $\delta = 0.20$ (the alternative hypothesis). At this stage, the researcher implausibly assumes that the underlying effect size is drawn from a distribution consisting of only two specific values. This is, of course, an unreasonably circumscribed distribution because the underlying effect size associated with the experiment in question could be literally anything, not just one of the two values briefly considered to compute statistical power.

Researchers who rely on Bayesian null hypothesis testing (BNHT) also typically assume that $\delta = 0$ under the null hypothesis (the point-null, as in NHST), but, under the alternative hypothesis, they do not assume that δ is a "point alternative" (such as $\delta = 0.20$). Instead, if the effect is real, they conceptualize δ as having been drawn from a continuous distribution, such as the Cauchy distribution (Rouder et al., 2009). An important – and, in our view, realistic – feature of this approach is that it assumes small effects are more likely than large effects even when the theory under investigation is true. As Rouder et al. (2009) put it: "One advantage of this setting is that small effects are assumed to occur with greater frequency than large ones, which is in accordance with what experimentalists tend to find" (p. 230). Even so, the zero/non-zero status of δ is still assumed to map directly on to the true/false status of the theoretical mechanism being tested. This is the key issue because the null hypothesis of no difference ($\delta = 0$) might never be strictly true. It might never be true because of what Baribault et al. (2018) referred to as "the vagaries and idiosyncrasies of experimental protocols" (p. 2607).

The critical but often overlooked distinction between the zero/non-zero status of δ and the true/false status of a theory under investigation was emphasized by Meehl (1967) long ago:

> While no competent psychologist is unaware of this obvious distinction between a substantive psychological theory T and a statistical hypothesis H implied by it, in practice there is a tendency to conflate the substantive theory with the statistical hypothesis, thereby illicitly conferring upon T somewhat the same degree of support given H by a successful refutation of the null hypothesis. (p. 107)

The problem with this way of thinking, as Meehl (1967) also pointed out, is that the null hypothesis of no difference is rarely if ever true *even if the theory under investigation is false*. In other words, the point-null hypothesis is a useful fiction that scientists adopt to perform statistical analyses, not a realistic depiction of underlying reality. Although $\delta = 0$ may never be strictly true, theories can be strictly false. This conundrum has long been appreciated by the field, but what it implies about the relationship between the magnitude of the underlying effect size and the true/false status of the theory under investigation has never been considered in a formal way, as we do here.

Our purpose is not to propose any new statistical test for dealing with the possibility that the null hypothesis of no difference is a fiction. Instead, we operate under the explicit assumption that, for the foreseeable future, analyses that adopt the point-null hypothesis – including both NHST and BNHT – will continue to dominate the statistical landscape. For the many studies that analyze their data using NHST or BNHT, our focus is on the relationship between the magnitude of δ (which could only be known if the entire population were tested) and the validity of a theoretical mechanism that correctly predicted its direction.

The goal of most methodological and statistical recommendations in the metascience literature is to ensure that a

claimed discovery is not a statistical false positive (i.e., to ensure that δ is not equal to zero). If the underlying effect size differs from zero and can be reproduced in large-*N* replication studies, the statistician is satisfied. This is true even if the effect size is small. It might not have practical relevance in that case – see, for example, efforts to specify the smallest effect size of interest (e.g., Lakens et al., 2018) – but so long as the underlying effect size is greater than zero, it is, by definition, a real effect, not a statistical illusion. But if the non-zero underlying effect size is small, is it also a theoretical true positive (i.e., does it also substantiate the theory that predicted it)?

With few exceptions, theories make predictions about the direction of an effect, not about the specific magnitude of an effect. Imagine that $\delta = 0.02$ for an experiment testing Theory A and that $\delta = 0.20$ for an experiment testing Theory B. Both underlying effect sizes differ from 0 in the direction predicted by the theory, so neither is a statistical false positive (Simmons et al., 2011), but are they equally likely to reflect true positives at the level of theory? We think that most scientists would say "yes." That is, a replicable non-zero effect supporting the a priori directional prediction made by a theory, no matter how small, is generally thought to support that theory to the same degree. However, we make the case here that the smaller the underlying effect size is, the less it supports the theory under investigation and the more likely it is to reflect a theoretically uninteresting nuisance factor.

As we use the term, a nuisance factor is any hidden cause of an underlying non-zero effect over and above the effect that might be contributed by a theoretical mechanism of interest. Examples of nuisance factors include (but are certainly not limited to) a pseudo-random number generator that leads to a slight confound across conditions; experimental instructions that unintentionally create a slight difference in attention to the task across conditions; research assistants who are not quite as blind to condition as the experimenter assumes; experimental participants who have taken a class related to the issue under investigation and manage to guess what the study is about; and so on. Regardless, if the effect caused by the nuisance factor is in the right direction, the theoretical mechanism under investigation will get the credit, even if the theory is wrong.[1]

The concept of a theoretical false positive is rarely considered, but the issue warrants attention because efforts to reduce false positives at the level of the underlying effect size, thereby mitigating the replication crisis, risk ushering in a new crisis at the level of theory. Indeed, one of the most widely agreed upon reforms for improving replicability is to increase sample size to enhance statistical power (e.g., Asendorpt et al., 2013; Bishop, 2019; Button et al., 2013; Turner et al., 2018). But if the null hypothesis of no difference is generally false, then, according to the argument we present here, this approach to increasing the detection of true positives at the level of the underlying effect size will have the unintended effect of increasing the detection of small-but-real effects that are false positives at the level of theory. Compounding that problem is the emerging idea that psychology should attach greater value to the discovery of small effects than it currently does (Götz et al., 2021).

Here, we advance the opposite perspective. In our view, a scientific discipline concerned with enhancing our understanding at the level of theory, like experimental psychology, should not become enamored with small effect sizes. In addition, we argue that there are limits to how much statistical power is advisable when an experiment is conducted to test a theory-based prediction. Such experiments fall into the category of what we term "original science," which is designed to expand the boundaries of knowledge (Wilson et al., 2020). Original science is distinct from "replication science," which is designed to confirm a candidate discovery from the original science literature by precisely measuring its underlying effect size.

As we discuss in more detail later, the rules that optimize original science are not the same rules that optimize replication science. For example, unlike theory-focused original science, where too much power can be problematic, for replication science, the higher the statistical power the better (Wilson et al., 2020). This is because all an experiment does is measure an effect size (nothing more). If an original experiment is among the small fraction of experiments deemed important enough to replicate, then the time has come to measure its underlying effect size as precisely as possible. Once the underlying effect size is precisely measured, further considerations are needed to decide if it is a theoretical true positive (reflecting a theoretical mechanism) or a theoretical false positive (reflecting a nuisance factor). One of those considerations is the magnitude of the precisely measured effect size.

In making our case, we emphasize two concepts that are likely to be unfamiliar to most researchers: (1) the concept of a *theoretical false positive* (as defined above) and (2) the concept of an *optimal* sample size (not too small but also not too large). The optimal sample size for original science is the one that maximizes positive predictive value (PPV) at the level of theory. PPV at the level of theory is the probability that a $p < .05$ result confirming a theory-based prediction reflects the effect of the theoretical mechanism, not a nuisance factor. Because our case rests on the notion that the null hypothesis of no difference is a useful fiction, not a realistic depiction of underlying reality, we next review the fascinating and enduring debate over that idea. Readers who need no convincing that the null hypothesis might never be strictly true can skip this section.

[1] Credit will be shared across competing theoretical mechanisms that happen to make the same prediction.

The null hypothesis of no difference is a fiction

Typically (though not necessarily), the null hypothesis states that the difference between the true underlying means of Condition A and Condition B (μ_A and μ_B, respectively) is *exactly* zero (i.e., μ_A - $\mu_B = 0$). In that case, the true underlying effect size would be exactly zero as well. For example, Cohen's $\delta = (\mu_A - \mu_B) / \sigma$, where σ is the common underlying standard deviation for both conditions (Cohen, 1988). In terms of this effect-size measure, the null hypothesis of no difference is that $\delta = 0$, and it is assumed to apply when the theoretical mechanism under investigation is false. This way of thinking implies that theoretically uninteresting nuisance factors do not create even so much as a small difference between groups when the theory is false. To us, and to many before us, this seems like an unrealistic assumption.

The way the underlying effect size is typically conceptualized if the theory is true (i.e., the alternative hypothesis) is similarly unrealistic. As noted earlier, when conducting a power analysis, researchers often assume that when the theory is true, the underlying effect size is a specific nonzero value (e.g., $\delta = 0.20$), as if no other possibilities exist. For Bayesians who use BNHT, with probability $P(H_0)$, the underlying effect size associated with the null hypothesis is also assumed to be $\delta = 0.00$ (as in NHST), but with probability $P(H_1)$, δ is not a point alternative. Instead, it is (more realistically) assumed to be drawn from a distribution, such as the Cauchy distribution.

But what if $\delta \neq 0.00$ even if the theoretical mechanism being tested is false? In that case, both statistical approaches would be based on a fictional (albeit useful) depiction of underlying reality. The idea that δ is never equal to zero is perhaps easiest to appreciate in quasi-experimental group designs in which the groups are either pre-existing (e.g., males vs. females, bilinguals vs. monolinguals, young vs. old, etc.) or created arbitrarily (e.g., those whose performance falls above the mean on some measure vs. those whose performance falls below the mean). Many experiments fall into this category. In the absence of random assignment, some difference between the two groups will always exist no matter what the dependent measure might be. Consider, for example, the results of a study by Bakan (1966):

> Some years ago, the author had occasion to run a number of tests of significance on a battery of tests collected on about 60,000 subjects from all over the United States. Every test came out significant. Dividing the cards by such arbitrary criteria as east versus west of the Mississippi River, Maine versus the rest of the country, North versus South, etc., all produced significant differences in means. (p. 425)

A similar issue emerges in studies that measure the correlation between one variable and another variable measured from the same participant. The issue here is that, within individuals, everything is correlated with everything else to some small degree. Therefore, with sufficiently large *N*, the result will be significant every time (Meehl, 1967; Vul et al., 2009). When participants are not randomly assigned to different experimental conditions, there appears to be no debate over the idea that the null hypothesis of no difference is always false despite what researchers pretend to be true when conducting statistical analyses (Meehl, 1990).

What about when participants *are* randomly assigned to different conditions? Here, there is room for argument, and the longstanding debate over this issue is as interesting as it is unresolved. Cohen (1990) was adamant that, yes, even then, the null hypothesis of no difference is always false:

> A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally . . . 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). (p. 1308)

Jones and Tukey (2000) emphatically agreed:

> When A and B are different treatments, μ_A and μ_B are certain to differ in some decimal place so that μ_A - $\mu_B = 0$ is known in advance to be false and μ_A - $\mu_B \neq 0$ is known to be true (Cohen, 1990; Tukey, 1991). An extensive rebuttal to this claim has been provided by Hagen (1997), who stated that "I agree that A and B will always produce differential effects on some variable or variables that theoretically could be measured. But I do not agree that A and B will always produce an effect on the dependent variable." (p. 20). We simply do not accept that view. For large, finite, treatment populations, a total census is at least conceivable, and we cannot imagine an outcome for which μ_A - $\mu_B = 0$ when the dependent variable (or any other variable) is measured to an indefinitely large number of decimal places. (p. 412)

However, a counterargument advanced by Hagen (1997) seems hard to summarily dismiss. He emphasized that arguments like the ones quoted above refer to *samples* taken from a population, yet the null hypothesis pertains to the population, not the samples. As he put it:

> But the null hypothesis says nothing about samples being equal, nor does the alternative hypothesis say that they are different. Rather, when addressing group differences, the null hypothesis says that the observed samples, given their differences, were drawn from the same population, and the alternative hypothesis says

Table 1 A summary of the distinction between theory-focused and measurement-focused research

- Theory-focused research
  - Goal is to test a prediction made by a theory
  - The binary decision (if $p < .05$ => credit the theory) is correct or incorrect
  - PPV at the level of theory (proportion correct) is maximized by optimizing N
- Measurement-focused research
  - Goal is to precisely measure the non-zero underlying effect size (δ)
  - Purely applied research is one type of measurement-focused research
  - When measuring δ, precision is maximized by maximizing N

> that they were drawn from different populations. (p. 20)

Hagen (1997) acknowledged that an experimental manipulation will always have an effect on *some* dependent measures despite random assignment to the treatment and control conditions (Conditions A and B, respectively). However, he also offered the following entertaining example to drive home the point that, for the specific dependent variable (DV) of interest, the null hypothesis of no difference is presumably true sometimes:

> A few years ago, visual imagery therapists were treating AIDS patients by asking the patients to imagine little AIDS viruses in their bodies being eaten by monsters…The effects of A and B are different, but many would question whether or not such changes would be reflected in the participant's T-cell count. And if that is the DV, it is only that difference that would lead to a rejection of the null hypothesis. (p. 21)

This argument seems reasonably compelling to us (see also Oakes, 1975). Then again, a stickler could argue that instructing patients to imagine virus-eating monsters in hopes of inducing phagocytes to ingest and digest HIV (the theoretical mechanism) might have some slight effect on adrenaline levels (the nuisance factor). Although a large effect of this experimental manipulation is hard to imagine, a small adrenaline-induced effect on T-cells is at least conceivable, one that would be detected with sufficient power. But no matter how real that small effect turns out to be, it would not lend much support to the proposed theoretical mechanism (phagocytosis).

Although some argue that small nuisance effects *always* exist (even when random assignment is used), it seems fair to suppose that there is some imaginable experimental manipulation that would have literally no effect on the dependent measure of interest, not even to the farthest decimal place imaginable (i.e., to infinity and beyond). Meehl (1990) himself eventually accepted the null hypothesis of no difference in "pure experimental studies" (p. 204) while also noting that his colleague David Lykken and "several high-caliber graduate students" disagreed with him on that point. We disagree with him on that point, too.

So far as we can determine, for experiments using random assignment, there has been no ultimate resolution to this debate. However, it is important not to overlook how narrow the focus of the remaining debate is. To our knowledge, no one disputes that the null hypothesis of no difference is always false in quasi-experimental and correlational studies. Moreover, few dispute that the null hypothesis of no difference is sometimes false even when random assignment is used. The only unresolved issue is whether the null hypothesis of no difference is *always* false when random assignment is used. Some say yes, others no.

Given how this debate has played out over the years, we assume that most scientists are willing to at least entertain the possibility that the null hypothesis of no difference is often not strictly true even though we pretend otherwise in virtually every statistical test we perform. What are the implications? We submit that the implications are non-trivial considering that theoreticians, unlike statisticians, often do not care about the magnitude of underlying effect sizes, per se.

Theory-focused versus measurement-focused research

Before presenting our case in formal terms, it is worth drawing another distinction between theory-focused research and measurement-focused research (Table 1). As noted above, theory-focused research investigates the truth status of a proposed theoretical mechanism.[2] Theories can be either true or false, even if underlying effect sizes associated with experiments designed to test them always differ from 0 to some degree. When the theory is false (e.g., people have

[2] Oberauer and Lewandowsky (2019) distinguished between *discovery-oriented* research and *theory-testing* research, which differ depending on whether the connection between the theory and the tested hypothesis is weak or strong, respectively. Both are subsumed by what we call theory-focused research.

Table 2 A summary of the distinction between original science and replication science

- Original science
  - Discovery-oriented research designed to expand the boundaries of knowledge
  - It includes testing a prediction derived from a theory (theory-focused research)
  - It also includes testing a purely applied question (measurement-focused research)
- Replication science
  - Confirmation-oriented research designed to measure the underlying effect size (δ) associated with a claimed discovery from original science
  - Replication science is measurement-focused and is optimized by maximizing *N*

ESP), we assume that the proposed theoretical mechanism (e.g., "quantum entanglement") contributes nothing to the non-zero underlying effect size associated with the experimental protocol. This is the null hypothesis at the level of theory, and, in our view, it is the null hypothesis of interest in theory-focused research. Later, we present a formal argument that PPV at the level of theory is maximized by optimizing (not maximizing) *N*.

Unlike theory-focused research, for measurement-focused research, the truth status of a proposed theoretical mechanism is not of interest. Instead, the question of interest is the magnitude of the underlying effect size per se (which is what statisticians focus on). Purely applied research is often measurement focused. Consider, for example, an experiment testing whether police lineups yield better performance when the faces are presented simultaneously or sequentially. Mickes et al. (2012) investigated that question after proposing what they argued was a better measure of diagnostic accuracy than had been used to that point, namely, the area under the receiver operating characteristic curve (AUC). At the time, no theory had been advanced to explain to why either lineup procedure would be diagnostically superior to the other according to AUC. Measurement was the only goal because a lineup procedure that achieves even slightly better diagnostic accuracy (e.g., slightly greater than the smallest effect size of interest), when multiplied across thousands of police departments, might be worth implementing in the real world. For measurement purposes, maximizing precision is achieved by maximizing *N*.

Now consider again the distinction between original science and replication science (Table 2). Original science is about expanding the boundaries of knowledge, whereas replication science is about confirming (or not) a finding from the original-science literature. Results obtained from experimental protocols in the original science literature do not directly answer the dichotomous questions of interest, such as (a) "Is the effect size different from 0 or not?" or (b) "Is the tested theoretical mechanism true or false?" Instead, because the measured effect size is a noisy estimate of the truth (i.e., it is a noisy estimate of the underlying effect size), a statistical test is needed to address such questions. At the replication stage, however, for the few original experiments that command attention, the goal should be to measure the truth as precisely as possible by maximizing *N*. In principle, though not always in practice, the measurement would be so precise as to render statistical analysis superfluous.

Imagine that a large-*N* replication study has provided a precise estimate of an underlying effect size of interest. If it is small yet undeniably greater than 0, it counts as a true positive at the level of the effect size (i.e., it is a statistical true positive). However, in what follows, we argue that the smaller the theory-supporting effect size is, the more likely it is to reflect a nuisance factor (i.e., the more likely it is to be a theoretical false positive). Theory-focused original-science research should try to avoid theoretical false positives and instead maximize the detection of theoretical true positives by optimizing, not maximizing, *N*.

In formal (but unrealistically simplified) terms

In a paper entitled "Power failure: Why small sample size undermines the reliability of neuroscience," Button et al. (2013) presented a seemingly airtight mathematical argument in favor of large-*N* studies. Their paper has been influential, having been cited over 6,000 times according to Google Scholar. Button et al. (2013) assumed that a reasonable scientific goal is to ensure that a high proportion of published $p < .05$ findings are statistical true positives. This proportion is known as positive predictive value, or PPV.[3] As they used this term, PPV refers to the proportion of published $p < .05$ findings in which δ is not equal to 0. This is PPV at the level of the underlying effect size, and many scientists implicitly assume that it is PPV at the level of theory as well (Meehl, 1967).

Adopting the Neyman and Pearson (1933) perspective, the equation specifying the relationship between *PPV* at the

[3] A more rational goal might be to maximize the ability to discriminate true hypotheses from false hypotheses (cf. Witt, 2019), but maximizing *PPV* is a sensible goal for a fixed alpha level. We proceed under the assumption that alpha is fixed at .05, in which case maximizing *PPV* is rational.

level of the underlying effect size, power ($1 - \beta$) and alpha (α) is:

$$\mathrm{PPV} = \frac{P(H_1)(1-\beta)}{P(H_1)(1-\beta) + P(H_0)\alpha} \quad (1)$$

where $P(H_1)$ is the prior probability that the alternative statistical hypothesis is true, and $P(H_0)$ is the prior probability that the null hypothesis of no difference is true. Because those are the only two possibilities, $P(H_0) + P(H_1) = 1$.

The prior *odds*, R, that the H_1 is true is given by $R = P(H_1)/P(H_0)$, so Equation 1 can be rewritten in the form used by Button et al. (2013):

$$\mathrm{PPV} = \frac{R(1-\beta)}{R(1-\beta) + \alpha} \quad (2)$$

One determinant of prior odds is how obvious the hypothesis being tested is in advance of the experiment. As a general rule, the more obvious it already is that H_1 is true, the higher the prior odds (in the limit, $R \to \infty$) and the more likely a significant effect is to replicate. Conversely, the less obvious it already is that H_1 is true, the lower the prior odds (in the limit, $R \to 0$), and the less likely a significant effect is to replicate. The appropriate level of R lies somewhere between these extremes, and it is a subjective judgment call. Different subfields of psychology (e.g., cognitive psychology vs. social psychology) may reasonably choose to operate at different points along that continuum (Wilson & Wixted, 2018). This means that the different subfields would have different replication rates even if they conducted methodologically similar experiments.

We assume that the prior odds for a given field are fixed, and for simplicity, we assume (as researchers typically do) that $P(H_0) = P(H_1) = 0.5$ such that $R = 1$ (i.e., the prior odds are even). If the prior odds are even, Equation 2 simplifies to:

$$\mathrm{PPV} = [(1-\beta)]/[(1-\beta) + \alpha] \quad (3)$$

From this perspective, science is a standard signal detection problem (Wilson et al., 2020). In terms of signal detection theory, power ($1 - \beta$) is the "hit rate" (*HR*) and α is the "false alarm rate" (*FAR*). Increasing N selectively increases the *HR* (i.e., it selectively increases power) while leaving the *FAR* fixed at $\alpha = .05$. Thus, as N increases, an ever-higher proportion of the $p < .05$ findings in the literature would be true positives at the level of the effect size (i.e., as N increases, so does PPV). If so, then assuming unlimited resources, only good things come from increasing N.

Critically, this argument is completely dependent on the assumption that under H_0, $\delta = 0$. For measurement-focused research, the null hypothesis of no difference is true by definition (i.e., by definition, H_0 means $\delta = 0$), and it does not matter that it may not apply to any experimental protocol. All that matters is the magnitude of the underlying effect size, whatever it might be, and an ever more precise measure of it is obtained by increasing N. However, our concern is with theory-focused research. Although δ may never equal zero because of nuisance factors, some theories are surely false in the strictest sense of that word. Unfortunately, as power is increased via large N to detect the non-zero effect predicted by a theory, power to detect the nuisance effect will increase as well. Thus, at the level of theory, both the *HR* and the *FAR* would increase as N increases. That is the problem.

Continuing with the assumption that the prior odds are even, consider the 50% of experiments that test a prediction made by a true theoretical mechanism. With sufficient power, these non-zero effects would all be detected at $p < .05$, and they would all be statistical true positives ($\delta > 0$ in the predicted direction). They would also all be theoretical true positives because the theory is true. For the 50% of experiments where the theory is false, the underlying effect caused by a nuisance factor would presumably be in the same direction predicted by the theory half the time (25% of the time overall) and in the opposite direction the other half of the time (again, 25% of the time overall). With sufficient power, these effects would also be detected, and they would all be statistical true positives because, in truth, $\delta \neq 0$. Unfortunately, when the nuisance effect happens to be in the same direction as that predicted by the theory (25% of the time overall), they would be theoretical false positives. Overall, 50% + 25% = 75% of experiments would confirm the theoretical prediction. Thus, with maximum power, PPV at the level of theory would be only .50 / (.50 + .25) = .67.

These considerations raise an important question: if the null hypothesis of no difference is never strictly true (as we and many other assume), then what would the relationship between PPV *at the level of theory* and N be? This turns out to be a more interesting question than it might seem at first glance.

For notational clarity, we redefine H_0 to represent the null hypothesis at the level of theory. Assume that under H_0 (the theoretical mechanism is false), the nuisance effect size is small but not zero. For example, assume that $\delta \,|\, H_0 = 0.02$ when the nuisance effect is in the same direction as the predicted effect, and $\delta \,|\, H_0 = -0.02$ when the nuisance effect is in the opposite direction as the predicted effect. Further assume that under H_1 (the theoretical mechanism is true), the effect size is 0.20 (i.e., $\delta \,|\, H_1 = 0.20$). To determine the quantitative relationship between PPV at the level of theory and N under those conditions, we used the appropriately modified version of Equation 3:

$$\text{PPV} = \left[\left(1-\beta_1\right)\right] / \left[\left(1-\beta_1\right)+\left(1-\beta_0\right)\right] \quad (4)$$

where $1-\beta_1$ is power to detect $\delta|H_1 = 0.20$ (as usual) and $1-\beta_0$ is power to detect $\delta|H_0 = 0.02$ (i.e., power to detect a directionally correct nuisance effect).[4] That is, we replaced α, which is constant with respect to N, with $1-\beta_0$, which increases with N.

For simplicity, we performed these calculations for a one-sample t-test, one-tailed (alpha level = .05), as if testing a directional effect predicted by a theoretical mechanism. With these settings, and as noted above, if power were 100%, 75% of experiments would yield a significant result in the predicted direction (50% because of the effect generated by a theoretical mechanism and 25% because a nuisance factor generated an effect in the predicted direction). In that case, as $N \rightarrow \infty$, PPV at the level of theory would be .50 / (.50 + .25) = .67. As illustrated in Fig. 1, for the parameter settings used in this example, the relationship between PPV at the level of theory and N is nonmonotonic, reaching a maximum of .941 at an intermediate sample size of $N = 284$. In an era where N can easily be in the thousands or tens of thousands (and sometimes even hundreds of thousands), Fig. 1 illustrates why a widely agreed upon reform to improve replicability has its limits.

With their "power failure" title, Button et al. (2013) cleverly implied that appliances and experiments alike need enough power to function properly. However, appliances and experiments alike can have too much power and therefore need a surge protector to guard against too much current. Optimizing rather than maximizing N provides a surge protector (so to speak) that will prevent an influx of theoretical false positives into the scientific literature.

In formal (and more realistic) terms

Although useful for illustrative purposes, the underlying effect size for a given experimental protocol is not realistically modeled as having been drawn from a point-null ($\delta \,|\, H_0 = 0.02$) distribution versus a point-alternative ($\delta \,|\, H_1 = 0.20$) distribution. Instead, the underlying effect size has been drawn from an unknown continuous distribution. What might that distribution look like across all experiments conducted by psychologists, whether or not the result is significant and whether or not it is published? There is no way to know. We can, however, try to take a principled approach to specify its possible shape.

As noted by Wilson et al. (2020), if all we know about a distribution is (1) its range and (2) its mean, then the *maximum entropy* distribution – that is, the distribution that is "…maximally noncommittal with regard to missing information" (Jaynes, 1957, p. 623) – is the exponential. With the direction of the underlying effect defined to be positive, the range is 0 to infinity. Although we do not know the exact mean of the distribution, we do have considerable information about it. For example, one estimated average effect size from the published social psychological literature is $\overline{d} = 0.43$ (Richard et al., 2003), which is likely inflated relative to the underlying effect sizes. It is also presumably larger than the underlying effect sizes of the many studies that were conducted and that were not published (e.g., perhaps because they failed to achieve statistical significance). Thus, out of the infinite range of possibilities, the mean underlying effect size presumably falls between 0 and 0.43.

Though not an exact estimate, after considering how much of the infinite range of possibilities we can exclude, Wilson et al. (2020) proceeded on the assumption that we know not only the range but also the mean. We therefore assumed that underlying effect sizes are drawn from an exponential distribution. Because that analysis was predicated on remaining maximally noncommittal about missing information, Wilson et al. (2020) stopped there. However, for thinking purposes, it seems reasonable to further suppose that, for some experiments, the theoretical mechanism under investigation is true (H_1) and for others, the theoretical mechanism is false (H_0). We therefore take that additional step here to work out the relationship between the underlying effect size (δ) and the likelihood that it was obtained in an experiment in which the theory under investigation is true. This is the relationship of primary interest. Later, we model

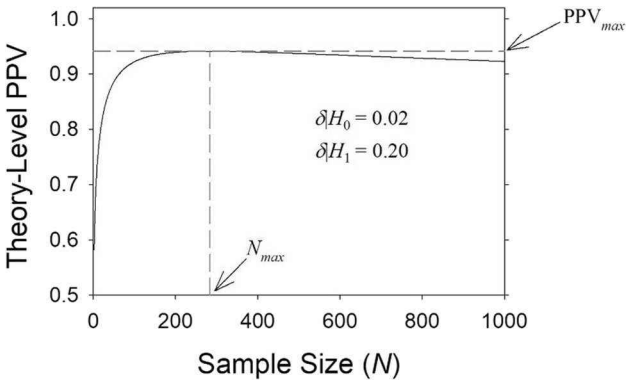

Fig. 1 Relationship between positive predictive value (PPV) at the level of theory vs. N under the assumption that $\delta|H_0 = 0.02$ and $\delta|H_1 = 0.20$ for a one-tailed one-sample t-test (with $\alpha = .05$). Under those conditions, the maximum PPV (PPV_{max}) of 0.941 is achieved using a sample size (N_{max}) of 284

[4] $1-\beta_0$ also includes power to falsely detect $\delta \,|\, H_0 = -0.02$ using a one-tailed t-test. These sign errors are rare and quickly become negligible as N increases.

the statistical selection of underlying effect sizes using a $p < .05$ filter for various N, where we again find that PPV at the level of theory is maximized when N is optimized, not when it is very small or very large.

Our model of science in a nutshell

A succinct one-paragraph summary of our model is as follows: If the theory that predicted the effect under investigation happens to be false, the non-zero underlying effect size associated with the experimental protocol arises from a nuisance factor alone. If the theory happens to be true, its underlying mechanism adds to whatever nuisance effect would exist even if the theory were false. By chance, the direction of the nuisance effect will be the same as the direction of the effect predicted by the theory half of the time (adding to the effect contributed by the theoretical mechanism, if the theory is true) and in the opposite direction the other half of the time (subtracting from the effect contributed by the theoretical mechanism, if the theory is true). Nuisance effects ("noise") and theory-based effects ("signal") are both assumed to be exponentially distributed, and an underlying effect size is conceptualized as a random draw from the noise distribution when the theory is false (half the time) and as a random draw from the signal-plus-noise distribution when the theory is true (the other half of the time).

Our model of science in detail

That is all there is to our model of underlying effect sizes. Everything we say next follows directly from it. We first work out the implications of this model for the relationship between the magnitude of δ for a given experimental protocol and the odds that the tested theory is true independent of any statistical test, and then we revisit the issue of PPV at the level of theory for statistically significant results. For readers who prefer to skip the math, Fig. 2 illustrates our model of underlying reality (under the assumption that nuisance effects are small relative to effects caused by a true theoretical mechanism), and Fig. 3 illustrates the corresponding relationship between the odds that the theory is true as a function of the magnitude of δ. Note that this relationship applies before any experimental result is selected using a $p < .05$ filter. PPV is a concept that applies after experimental results are selected in this way, and Fig. 6a depicts PPV at the level of theory as a function of N for the model shown in Fig. 2. Figures 4, 5 and 6b provide another example under the assumption that nuisance effects are, on average, as large as the effects generated by true theoretical mechanisms. The take-home message is the same either way: the smaller the non-zero magnitude of δ, the less support it offers for the theoretical mechanism that predicted it and the more likely it is to reflect a nuisance factor. In addition, PPV is maximized by optimizing (not maximizing) N.

Model mechanics We begin by partitioning δ into two independent components, δ_N and δ_S, where the subscript "N" stands for "noise (the share of the effect caused by a nuisance factor), and the subscript "S" stands for "signal" (the share of the effect caused by the theoretical mechanism under investigation). If the theory that predicted the effect is false, then $\delta_S = 0$. This is the null hypothesis at the level of theory (H_0). In that case, $\delta = \delta_N$. When the theory is true, then $\delta_S > 0$, in which case $\delta = \delta_S + \delta_N = \delta_{SN}$, where the subscript "$SN$" stands for "signal plus noise." This is the alternative hypothesis at the level of theory (H_1). These assumptions are summarized in Table 3.

When the theory being tested is true, we assume that δ_S is a random variable (x), falling in the positive range of $(0, \infty)$. Thus, by definition, the direction of the effect predicted by the theory is positive. We take this unidirectional approach because it is hard to imagine a useful theory that predicts a non-zero effect but does not predict its direction (Jones & Tukey, 2000). More formally, we assume that x is an exponentially distributed random variable with rate parameter λ, the pdf of which is:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Figure 2a illustrates this signal distribution for $\lambda = 5$ (mean effect size $= 1 / \lambda = .20$).

Unlike δ_S, we assume that δ_N always differs from 0. After all, there is only one theory under investigation in the simplest case (which might be false), but there is an inexhaustible supply of possible nuisance factors that might affect the dependent measure for a given experimental protocol. Even a standard random number generator does not generate truly random numbers but instead generates pseudo-random numbers that must introduce some (perhaps infinitesimal) non-zero effect. We therefore assume that one or more nuisance factors plagues *every* experimental protocol. However, the essence of our argument (i.e., that theoretical false positives will increase as power to detect nuisance effects increases) holds even if the null hypothesis of absolute 0 is sometimes true. Only the math would change, as we consider later. The argument is undermined only if virtually every experimental protocol is so pristine as to be entirely free of nuisance factors, in which case there would be little need to ever worry about potential non-obvious confounds.

Formally, we assume that δ_N is a bidirectional, exponentially distributed random variable (x) with rate parameter μ, and its direction is in the same direction as that predicted by the theory ($x > 0$) half the time and in the opposite direction

($x<0$) the other half of the time. Thus, the pdf of the noise distribution is:

$$g(x,\mu)=\begin{cases}0.5\mu e^{-\mu x} & \text{if } x\geq 0 \text{ (same direction)}\\ 0.5\mu e^{\mu x} & \text{if } x\leq 0 \text{ (opposite direction)}\end{cases} \quad (5)$$

This can also be written more succinctly as $g(x,\mu)=0.5\mu e^{-\mu|x|}$.The mean of this mirror-imaged exponential distribution is $1/\mu$ for $x>0$ and $-1/\mu$ for $x<0$, so its overall mean is $\frac{1/\mu+(-1/\mu)}{2}=0$.[5] Figure 2b illustrates these pure "noise" experiments with $\mu=50$ (i.e., for positive x, the mean equals $1/\mu=.02$). Note that with a mean this small, many underlying effect sizes are negligible (e.g., ~5% are less than $\delta_N=.001$ and 1% are less than $\delta_N=.0002$). However, none are so small as to equal 0 exactly.

With both the signal and noise distributions defined, we can now specify the pdf of the signal-plus-noise distribution (i.e., the distribution of δ_{SN}), where $\delta_{SN}=\delta_S+\delta_N$. This is the distribution of underlying effect sizes when the theory under investigation is true. Recall that when the tested theory is true, $\delta_S>0$, which is to say that its direction is positive (by definition). In that case, if δ_N happens to be in the same direction as δ_S (i.e., both positive), then their summed value is necessarily positive as well ($\delta_{SN}>0$). However, if δ_N is in the opposite direction, as it will be half the time, then the summed value can still be positive (when $\delta_S>\delta_N$), but it can also be negative (when $\delta_S<\delta_N$). As detailed in the Appendix, for the general case of $\lambda\neq\mu$, δ_{SN} is a random variable (x) distributed as follows:

$$h(x,\lambda,\mu)=\begin{cases}0.5a\left(e^{-\mu x}-e^{-\lambda x}\right)+0.5b\left(e^{-\lambda x}\right) & \text{if } x\geq 0\\ 0.5b(e^{\mu x}) & \text{if } x\leq 0\end{cases} \quad (6)$$

where $a=\frac{\lambda\mu}{\lambda-\mu}$ and $b=\frac{\lambda\mu}{\lambda+\mu}$. For $\lambda=5$ and $\mu=50$, the signal-plus-noise distribution – that is, the distribution of δ_{SN} – is the one shown in Fig. 2c.

Figure 2c indicates that when the theoretical mechanism under investigation is true, the underlying effect size (i.e., the effect size that would be measured if the entire population were tested) is usually in the direction predicted by the theory. However, occasionally, the underlying effect is in the opposite direction predicted by a true theory. This happens when the effect of an opposite-direction nuisance factor more than cancels the effect contributed by the theoretical mechanism.

The likelihood ratio Having specified both the noise distribution (the distribution of δ_N) and the signal-plus-noise distribution (the distribution of δ_{SN}), we are now in a position to compute the likelihood that the effect is due to a theoretical mechanism (compared to the likelihood that the effect is instead due to a nuisance factor) as a function of the magnitude of δ. As a reminder, for half the experiments, we conceptualize δ as a random draw from the noise distribution (Fig. 2b) and, for the other half, as a random draw from the signal-plus-noise distribution (Fig. 2c).

For any given underlying effect size $\delta=x_i$, there is some probability that it is a signal (plus noise) trial, $P(x_i|s)$, and some probability that it is instead a noise trial, $P(x_i|n)$. Using the functions presented above, $P(x_i|s)=h(x_i,\lambda,\mu)$ and $P(x_i|n)=g(x_i,\mu)$.[6] The likelihood ratio, $L(x_i)$, is equal to $L(x_i)=P(x_i|s)/P(x_i|n)$. Thus, for $x=\delta>0$ (i.e., when the underlying effect size is in the direction predicted by the theory), the likelihood ratio is:

$$L\left(x_i|x_i>0\right)=\frac{0.5a\left(e^{-\mu x_i}-e^{-\lambda x_i}\right)+0.5b\left(e^{-\lambda x_i}\right)}{0.5\mu e^{-\mu x_i}}$$

With some algebraic rearrangement, this expression can be written as:

$$L\left(x_i|x>0\right)=\left(k_1+k_2\right)e^{(\mu-\lambda)x_i}-k_1 \quad (7)$$

where $k_1=\frac{\lambda}{\mu-\lambda}$ and $k_2=\frac{\lambda}{\mu+\lambda}$.[7]

Equation 7 states that the likelihood ratio increases as a function of x. In other words, the larger the underlying effect size, the more support it offers for the theoretical mechanism under investigation. Although it seems reasonable to suppose that nuisance effects are relatively small, this relationship holds true no matter what the non-negative values of λ and μ might be. Thus, our argument is not dependent on the assumption that effects generated by nuisance factors are smaller than effects generated by the theoretical mechanism of interest. This can be appreciated by considering the first derivative of Equation 7. Dropping the subscript i on x, the first derivative of the likelihood ratio is:

$$\frac{d}{dx}\left[\left(k_1+k_2\right)e^{(\mu-\lambda)x}-k_1\right]=\frac{2\lambda\mu}{\mu+\lambda}e^{(\mu-\lambda)x} \quad (8)$$

The derivative on the right is the slope of the likelihood ratio versus x (where x represents the magnitude of δ). Because $e^{(\mu-\lambda)x}$ is positive for $x>0$, the slope of the function

[5] Equation 5 is the Laplace distribution with a mean and mode of 0. However, this does not mean that an underlying effect size of 0 ever occurs. Instead, it means that an underlying effect size is more likely to be close to 0 than to any other value.

[6] $P(x_i|s)$ means "probability of x_i given that signal is present." However, because noise is always present as well, this probability is provided by the signal-plus-noise distribution.

[7] Again, this equation is for the general case where $\lambda\neq\mu$. See Appendix for the corresponding equation for the special case where $\lambda=\mu$ (the equation itself is different, but the implications are the same).

Table 3 The magnitude of the underlying effect size (δ) is conceptualized as the sum of an effect generated by a theoretical mechanism (δ_S) plus an effect generated by a nuisance factor (δ_N)

- Signal (δ_S)
  - The share of δ caused by the operation of a theoretical mechanism
  - If the theoretical mechanism is true, $\delta_S > 0$; if it is false, $\delta_S = 0$
- Noise (δ_N)
  - The share of δ caused by an undetected nuisance factor
  - $\delta_N \neq 0$ and is either positive or negative with respect to the theory-based prediction
- Signal + Noise ($\delta_S + \delta_N$)
  - For a given experimental protocol, $\delta = \delta_S + \delta_N$
  - If the theoretical mechanism is false ($\delta_S = 0$), $\delta = 0 + \delta_N = \delta_N$
  - If the theoretical mechanism is true ($\delta_S > 0$), $\delta = \delta_S + \delta_N = \delta_{SN}$

relating the likelihood ratio to x is also positive for all $x > 0$. This means that the odds that the effect reflects the operation of a theoretical mechanism increases with x, and the key point is that this is true no matter what (positive) λ and μ might be. If $\mu > \lambda$ (i.e., if nuisance effects are smaller than theory-driven effects, on average), the derivative eventually explodes to infinity as x increases. If $\mu < \lambda$ (i.e., if nuisance effects are larger than theory-driven effects, on average), the derivative eventually asymptotes at 0 as x increases. But the slope is always positive, which means that the likelihood ratio always increases with x. This, in turn, means that the larger the underlying effect size, the more likely it is to reflect the operation of a theoretical mechanism, not a nuisance factor.

Consider next how our model of underlying reality relates to the almost universal assumption that the null hypothesis of no difference is strictly true when the theory is false (i.e., when the theory is false, $\delta = 0$). As μ increases, the mean of the noise distribution for positive x decreases (i.e., as $\mu \to \infty$, $\overline{\delta_N} \to 0$), at which point the null hypothesis of no difference is true even at the level of theory. As $\mu \to \infty$, the likelihood ratio in Equation 7 approaches ∞ even for $x = 0 + \epsilon$ (i.e., even for x slightly greater than 0). In other words, *any* nonzero underlying effect size, no matter how small, indicates that the effect is due to the theoretical mechanism that predicted it. Thus, as $\mu \to \infty$, the null hypothesis at the level of theory reduces to the standard point-null hypothesis.

Figure 3a illustrates the relationship between the log likelihood ratio (i.e., log odds) and the underlying effect size for $\lambda = 5$ ($\overline{\delta_S} = 0.20$) and $\mu = 50$ ($\overline{\delta_N} = 0.02$ for $x > 0$), and Fig. 3b presents the same information expressed as a probability.

If $L(x_i | x > 0) > 1$, then the underlying effect size is more likely to reflect the hypothesized theoretical mechanism (signal plus noise) than it is to reflect a nuisance factor (noise alone). If $L(x_i | x > 0) < 1$, it is the other way around. And if $L(x_i | x > 0) = 1$, the odds are even. For convenience, we consider the log transform of the likelihood ratio in Fig. 3a. Negative log likelihood ratios mean that the underlying effect size likely reflects a nuisance factor, whereas positive log likelihood ratios mean that the underlying effect size likely reflects the hypothesized theoretical mechanism. The log likelihood ratio is equal to zero (and the probability equals .50) for the underlying effect size

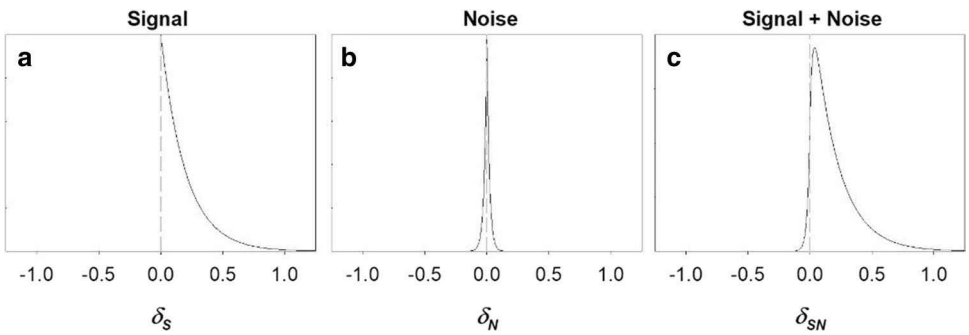

Fig. 2 **a** Exponential distribution of effect sizes generated by the theoretical mechanism of interest when the theory is true (δ_S). The mean of this distribution is 0.20. The direction of the effect predicted by the theory is defined to be positive, so negative values of δ_S do not exist. **b** Bidirectional exponential distribution of effect sizes generated by nuisance factors whether the theory is true or false (δ_N). The mean is equal to 0.02 when the nuisance effect is in the same direction predicted by the theory and is equal to -0.02 when it is in the opposite direction. **c** Distribution of the sum of δ_S and δ_N (signal plus noise, or δ_{SN}) when the theory under investigation is true

where the odds are even (which occurs at ~0.04 in this example).

The key point is that the smaller the underlying effect size, the less support it provides for the theory despite being a true (i.e., non-zero) effect in the direction predicted by the theory. This is precisely why, for theory-focused research, maximizing N to enhance replicability eventually becomes counterproductive. Despite maximizing replicability, it becomes counterproductive because ever-smaller values of δ would be detected at $p < .05$ and appear in the scientific literature as new theoretical discoveries when they actually reflect nuisance factors.

Another example To this point, we have used an example where nuisance effects are small, on average ($\mu = 50$, so the mean of δ_N for $x > 0$ is equal to $1 / 50 = 0.02$), relative to theory-focused effects ($\lambda = 5$, so the mean of δ_S is equal to $1 / 5 = 0.20$). We used this asymmetrical example because it is not a dramatic departure from what scientists already implicitly assume, which is that $\delta_N = 0$ (the point-null hypothesis), with δ_S being some value substantially greater than that. However, there is no way to know if this intuition-based asymmetry in the relative size of signal versus noise effects is correct. To remain maximally noncommittal to unknown information, one might assume that nuisance effects and theory-based effects are, on average, equivalent. As detailed in the Appendix, in that special case (i.e., when $\lambda = \mu$), the signal-plus-noise distribution becomes:

$$h(x, \mu) = \begin{cases} 0.5\mu e^{-\mu x}(\mu x + 0.5) & \text{if } x \geq 0 \\ 0.25\mu e^{\mu x} & \text{if } x \leq 0 \end{cases}$$

The noise distribution is the same as before:

$$g(x, \mu) = \begin{cases} 0.5\mu e^{-\mu x} & \text{if } x \geq 0 \text{ (samedirection)} \\ 0.5\mu e^{\mu x} & \text{if } x \leq 0 \text{ (opposite direction)} \end{cases}$$

As an example, assume that $\lambda = \mu = 5$. In that case, the mean underlying effect size for nuisance factors and the mean underlying effect size contributed by true theoretical mechanisms would both be $\mu = 0.20$. Figure 4 illustrates the signal distribution (which is the same as before), the noise distribution, and the signal-plus-noise distribution for this scenario.

Figure 5a illustrates the corresponding relationship between mean underlying effect size and the log odds that the theory is true, and Fig. 5b presents the same information in terms of probability. Even if, on average, nuisance effects are this large, it nevertheless remains true that ever smaller underlying effect sizes for a given experimental protocol offer less support for the theory that predicted it. What changes is the slope of the function, which is shallower than it was before. Now, the odds are even when the underlying effect size is 0.10, and the odds favor a nuisance factor when the underlying effect size is smaller than that.

As it turns out, and as is evident from the first derivative of the likelihood ratio function presented earlier in Equation 8, the key principle remains the same no matter what the exact positive values of λ and μ might be: larger underlying effects are more supportive of the theory that predicted them than smaller underlying effects (and vice versa). Yet to address the replication crisis, increasing N without bound is increasingly considered

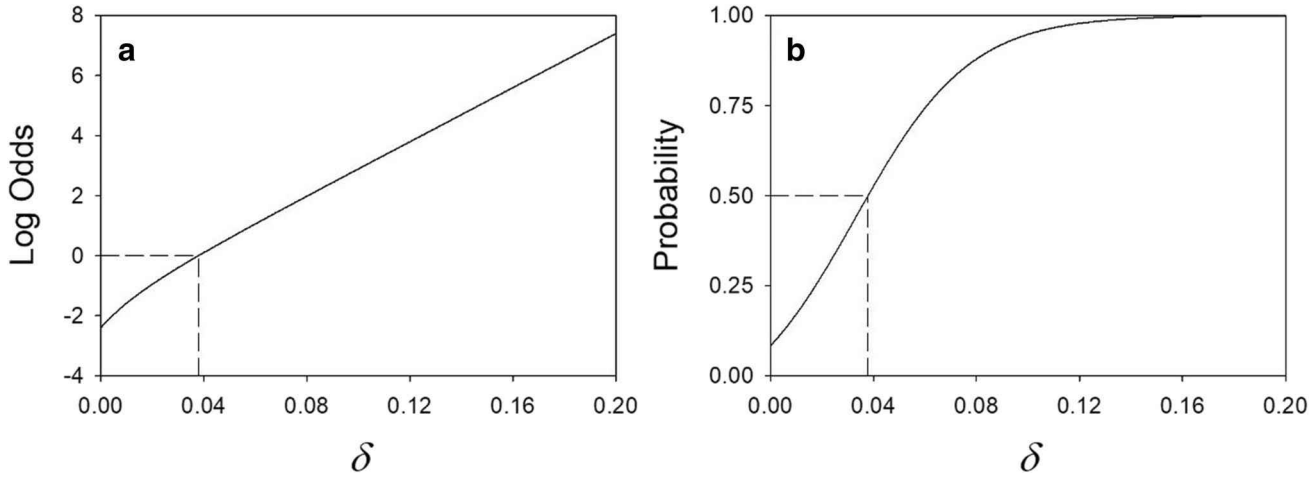

Fig. 3 **a** Log odds that the theory of interest is true as a function of the magnitude of the underlying effect size (δ) for the model depicted in Fig. 2. For half the experiments, the theory was assumed to be false (so the effect size was a random draw from Fig. 2b) and for the other half it was assumed to be true (so the effect size was a random draw from Fig. 2c). The dashed lines indicate the underlying non-zero effect size for which the odds are even that the effect is due to a nuisance factor vs. the theoretical mechanism of interest. **b** The same information expressed in terms of the probability that the theory of interest is true

to provide only a benefit (increasing replicability, which it will certainly do), without cost beyond the increased consumption of resources. This is true for measurement-focused research, but for theory-focused research, there is another cost that seems too high to pay. The cost is the increased introduction of small underlying effect sizes into the scientific literature, which will often be theoretical false positives masquerading as true theoretical discoveries.

What if the null hypothesis of no difference is sometimes true? We have assumed that nuisance effects always exist, drawn from an exponential distribution. However, allowing for the possibility that some experimental protocols achieve absolute methodological perfection (in which case the null hypothesis of no difference is literally true to the infinite decimal place) would not change our conclusion. For example, assume three possible outcomes for the noise distribution (the theory is false):

$$g(x,\mu)=\begin{cases}(1-\pi)0.5\mu e^{-\mu x} & \text{if } x>0 \text{ (same direction)}\\ (\pi)0 & \text{if } x=0\\ (1-\pi)0.5\mu e^{\mu x} & \text{if } x<0 \text{ (opposite direction)}\end{cases}$$

where π represents the proportion of experiments in which the null hypothesis of no difference is strictly true (i.e., the proportion of experiments where $\delta=0$). In that case, the probability of a nuisance factor generating an effect in the positive direction would be $(1-\pi)0.5$, as would the probability of a nuisance factor generating an effect in the opposite direction. For the general case where $\lambda\neq\mu$, the likelihood ratio for $x>0$ now becomes:

$$L(x_i|x>0)=\frac{0.5a\left(e^{-\mu x_i}-e^{-\lambda x_i}\right)+0.5b\left(e^{-\lambda x_i}\right)}{(\pi)0+(1-\pi)0.5\mu e^{-\mu x_i}}$$

where the term $(\pi)0$ in the denominator is included for the sake of clarity even though it equals 0. With some algebraic rearrangement as before, this expression can be written as:

$$L(x_i|x>0)=\frac{\left(k_1+k_2\right)e^{(\mu-\lambda)x_i}-k_1}{(1-\pi)} \tag{9}$$

Although we have assumed throughout that $\pi=0$ (i.e., the null hypothesis of no difference is never strictly true even when the theoretical mechanism is false), Equation 9 shows that our story does not depend on that assumption. So long as $\pi<1$ (i.e., if the null hypothesis of no difference is sometimes false when the theoretical mechanism under investigation is false), the take-home message is the same: the smaller the underlying effect size, the less likely it is that the theoretical mechanism that predicted it is true. It is only when $\pi=1$ (i.e., the null hypothesis of no difference is always strictly true when the theoretical mechanism is false) that the likelihood ratio becomes infinite for any non-zero underlying effect size, no matter how small. For theory-focused research, scientists who rely on standard NHST or BNHT adopt that assumption implicitly.

Positive predictive value (PPV) at the level of theory

So far, we have worked out the relationship between the underlying effect size associated with an experimental protocol (which could be known only if the entire population were tested) and the probability that it reflects the theoretical mechanism that predicted it rather than a nuisance factor. Because small effects are likely to reflect nuisance factors, the implication is that routinely maximizing *N* would

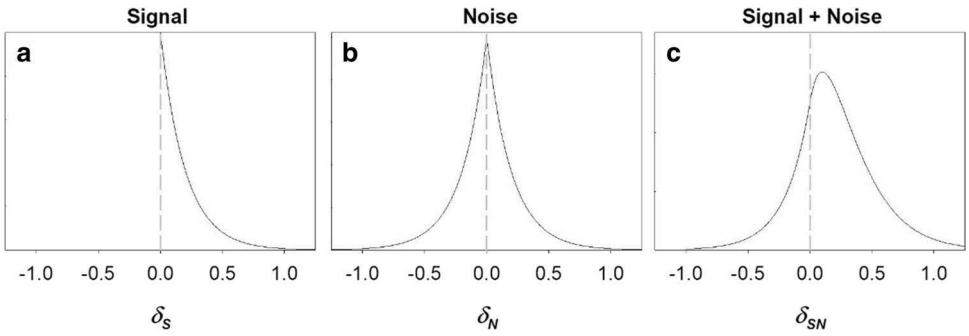

Fig. 4 **a** Exponential distribution of effect sizes generated by a true theoretical mechanism (δ_S). The mean of this distribution is 0.20. **b** Exponential distribution of effect sizes generated by nuisance factors whether the theory is true or false (δ_N). The mean is now equal to 0.20 for positive δ_N and is equal to -0.20 for negative δ_N. **c** Distribution of the sum of δ_S and δ_N (signal plus noise, or δ_{SN}) when the theory under investigation is true. When nuisance effects are larger than effects caused by the theoretical mechanism of interest, underlying effect sizes opposite to the theoretically-predicted direction are fairly common (25% of the time in this example) even when the theory is true

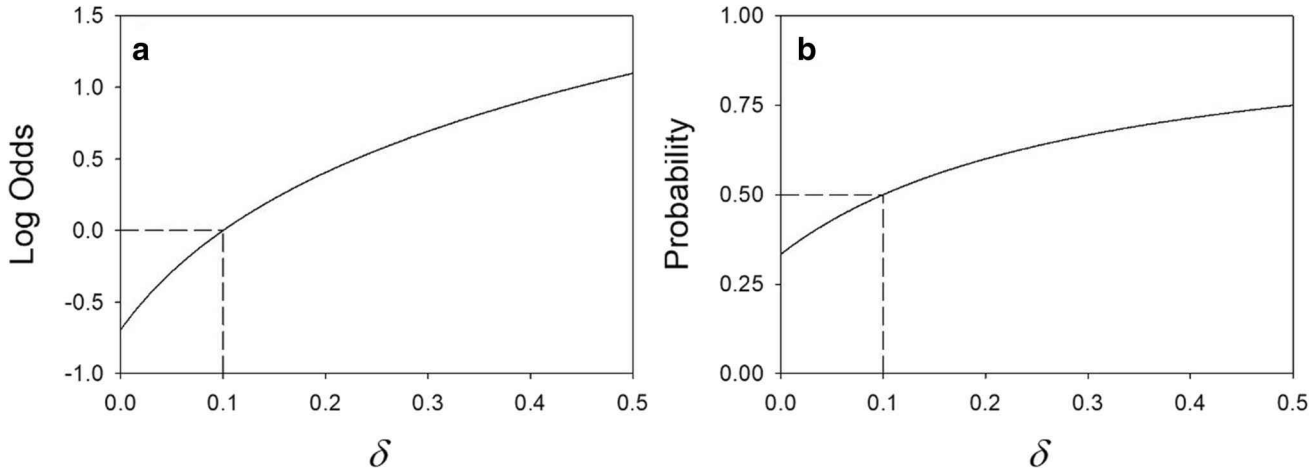

Fig. 5 **a** Log odds that the theory of interest is true as a function of the magnitude of the underlying effect size (δ) for the scenario depicted in Fig. 4. This function was generated assuming that for half the experiments, the theory was false (so the effect size was a random draw from Fig. 4b) and for the other half of the experiments, the theory was true (so the effect size was a random draw from Fig. 4c). The dashed lines indicate the underlying non-zero effect size for which the odds are even that the effect is due to a nuisance factor vs. the theoretical mechanism of interest. **b** The same information expressed in terms of the probability that the theory of interest is true

be problematic because small nuisance effects would be routinely detected at $p < .05$. Should we therefore minimize N? That would be an even worse idea.

Figure 6 shows PPV at the level of theory for statistically significant ($p < .05$) results as a function of N for the two models of underlying reality considered above (Figs. 2 and 4). Figure 6a assumes the model of underlying effect sizes depicted in Fig. 2 such that $\overline{\delta_S} = .20$ and $\overline{\delta_N} = .02$ for $x > 0$. Figure 6b assumes the model of underlying effect sizes depicted in Fig. 4 such that $\overline{\delta_S} = .20$ and $\overline{\delta_N} = .20$ for $x > 0$. The statistical analyses testing for $p < .05$ effects involved one-tailed, one-sample t-tests (see Appendix for mathematical details). Clearly, as we also found earlier when using a much simpler model of underlying reality, PPV at the level of theory is maximized by neither minimizing nor maximizing N but is instead maximized using an intermediate, optimal sample size (N_{max}).

The optimal N in Fig. 6a is $N_{max} = 242$. It is optimal in the sense that it maximizes PPV at the level of theory for the scenario in which $\overline{\delta_S} = .20$ and $\overline{\delta_N} = .02$ and $x > 0$. As the mean of the noise distribution increases,

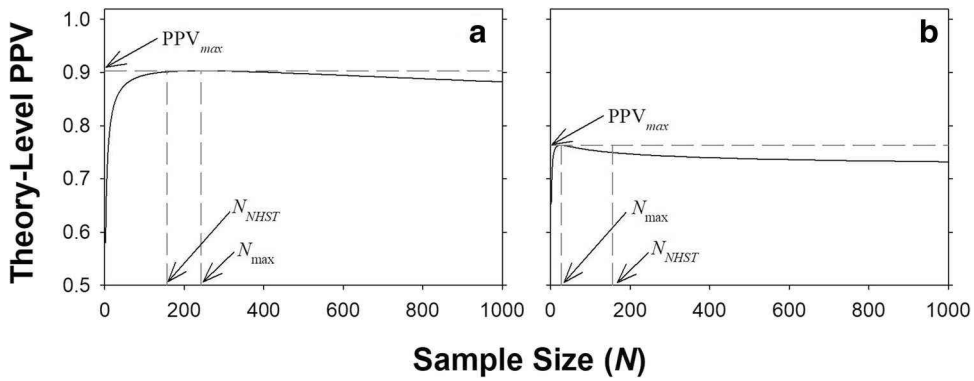

Fig. 6 **a** Positive predictive value (PPV) at the level of theory as a function of sample size (N) for the model depicted in Fig. 2 ($\overline{\delta_S} = .20$, $\overline{\delta_N} = .02$ for positive x). The maximum achievable PPV (PPV_{max}) is equal to .903, and the sample size that achieves that value (N_{max}) is 242. **b** PPV at the level of theory as a function of sample size (N) for the model depicted in Fig. 4 ($\overline{\delta_S} = .20$, $\overline{\delta_N} = .20$ for positive x). The maximum achievable PPV (PPV_{max}) is equal to .763, and the sample size that achieves that value (N_{max}) is 27. Also shown is the sample size that one would use to achieve 80% power (N_{NHST}) assuming that $\delta = 0$ under the null hypothesis and $\delta = 0.20$ under the alternative hypothesis. For a one-tailed, one-sample t-test under those conditions, $N_{NHST} = 156$ (its value is the same in panels **a** and **b**)

however, optimal N decreases rapidly. In Fig. 6b, where $\overline{\delta_S} = \overline{\delta_N} = .20$ for $x > 0$, optimal $N_{max} = 27$.

In Fig. 6a, PPV is, for all intents and purposes, maximized using the N that one would use to achieve 80% power to detect a small underlying effect size ($\delta = 0.20$) under the standard assumptions of NHST (i.e., $\delta = 0$ under the null hypothesis). In this scenario, based on standard NHST logic, 80% power would be achieved be testing $N_{NHST} = 156$ participants. By most accounts, this would be an adequately powered study. By contrast, in Fig. 6b, where the mean of the noise distribution is equal to the mean of the signal distribution, the use of $N_{NHST} = 156$ participants would result in an overpowered experiment.

The key point is that, according to either version of underlying reality (Fig. 2 or Fig. 4), using large values of N reduces PPV in addition to consuming more resources. On the positive side, it also provides a more accurate estimate of δ. However, that benefit quickly yields diminishing returns. Figure 7 illustrates the average observed effect size (Cohen's d) and the average underlying effect size (Cohen's δ) as a function of N for statistically significant ($p < .05$) results. Figure 7a assumes the model shown in Fig. 2 ($\overline{\delta_S} = .20$ and $\overline{\delta_N} = .02$ for $x > 0$), and Fig. 7b assumes the model shown in Fig. 4 ($\overline{\delta_S} = .20$ and $\overline{\delta_N} = .20$ for $x > 0$). The mathematical formulas used to generate these plots are presented in the Appendix.

Note how inflated the observed $p < .05$ effect size is for small N (dashed curves). However, for $N_{NHST} = 156$, Cohen's d already provides a reasonably good estimate of Cohen's δ, on average. Increasing N beyond that would further reduce the confidence interval around Cohen's d, but that minor benefit would come at the major cost of detecting ever more theoretical false positives masquerading as theoretical true positives.

Figure 7 illustrates another point that we have not considered thus far here but did consider in Wilson et al. (2020). Whereas Fig. 6 illustrates that PPV is maximized using the optimal sample size of N_{max}, Fig. 7 illustrates that the *mean* of the underlying $p < .05$ effect size (i.e., $\overline{\delta} \mid p < .05$) is maximized using a much smaller N. More specifically, $\overline{\delta}_{max}$ is achieved with $N = 24$ and $N = 12$ in Fig. 7a and b, respectively. Why care about that measure? Without any dichotomous consideration of effect sizes arising from theoretical mechanisms versus nuisance factors, Wilson et al. (2020) assumed that underlying effect sizes were drawn from a single exponential distribution. A model of underlying reality like that would apply if, when testing non-obvious predictions made by a theory, by assumption (not by mathematical derivation), the larger the underlying effect size, the more likely it is to reflect the theoretical mechanism rather than a nuisance factor. This is a continuous version of the dichotomous model of underlying reality that we have pursued here. In the continuous version that Wilson et al. (2020) considered, the concept of PPV is not directly quantifiable. Given that model of underlying reality, powering experiments to maximize $\overline{\delta} \mid p < .05$ would be a rational goal.

Yet scientists routinely assume, implicitly or explicitly, that half the tested theory-based hypotheses are true, and half are false (e.g., Rouder et al., 2009). After adopting that assumption here in a model of exponentially distributed underlying effect sizes, another optimal value of N is the one that maximizes PPV at the level of theory (Fig. 6). When nuisance effects are relatively small, the sample size that maximizes PPV at the level of theory is quite a bit larger than the sample size that maximizes the mean of $p < .05$ underlying effect sizes. Striving to maximize PPV at the level of theory seems like a reasonable goal for theory-focused research so long as one is willing to grant the

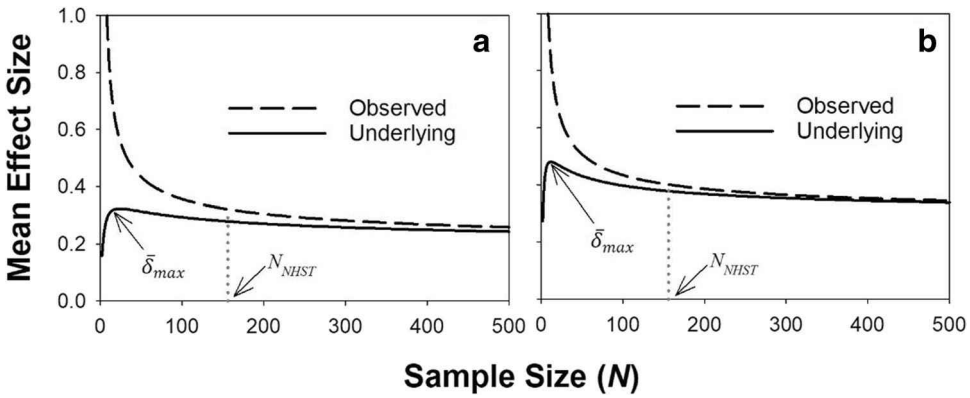

Fig. 7 **a** Mean observed effect size and underlying effect size (given $p < .05$) as a function of sample size (N) for the model depicted in Fig. 2 ($\overline{\delta_S} = .20$, $\overline{\delta_N} = .02$ for positive x). **b** Mean observed and underlying effect size (given $p < .05$) as a function of sample size (N) for the model depicted in Fig. 4 ($\overline{\delta_S} = .20$, $\overline{\delta_N} = .20$ for positive x). In panels **a** and **b**, $\overline{\delta}_{max}$ represents the maximum of $\overline{\delta} \mid p < .05$ (i.e., the maximum of the mean δ for statistically significant results). As in Fig. 6, N_{NHST} is the sample size that one would use to achieve 80% power assuming $\delta = 0$ under the null hypothesis and $\delta = 0.20$ under the alternative hypothesis for a one-tailed, one-sample t-test ($N_{NHST} = 156$ in panels **a** and **b**)

assumptions of the dichotomous model illustrated earlier in Figs. 2 and 4. But whether the goal is to maximize $\overline{\delta}$ for $p < .05$ results (Fig. 7) or to maximize PPV at the level of theory (Fig. 6), it is achieved by optimizing rather than maximizing *N*.

As noted above, and as illustrated in Fig. 7, when *N* is small and power is correspondingly low, the expected value of a $p < .05$ Cohen's *d* is (on average) massively inflated. Thus, if experimental psychologists often run underpowered studies, as they undoubtedly do, replication studies (which are not selected using a $p < .05$ filter and therefore do not yield inflated estimates of Cohen's *d*, on average) will have substantially smaller effect sizes than the original studies simply due to regression to the mean. This can create the impression of a replication crisis when the truth is considerably more nuanced than that (Maxwell et al., 2015).

The replication crisis reconsidered

The debate over the existence of and proposed solutions to the replication crisis has focused almost exclusively on measured effect sizes, as if all of science consists of measurement-focused research. Consider the Open Science Collaboration (2015, hereafter OSC2015), which directly replicated 100 representative original-science experiments from the psychology literature, 97 of which reported a significant result. For those 97, the replication effect sizes were smaller than the original findings that were selected using a $p < .05$ filter, and less than 40% of the replication experiments again yielded a $p < .05$ result.

If over 60% of the original studies replicated in OSC2015 were statistical false positives (i.e., $\delta = 0$), as many assume, then their corresponding replication Cohen's *d* effect sizes would be centered on 0. However, Wilson et al. (2020) analyzed the non-significant replications and found that the average effect size was $\overline{d} = 0.141$, an outcome significantly greater than 0 at $p < .001$. Similarly, a Bayesian analysis performed on these data yields a Bayes factor that strongly favors the alternative hypothesis over the point-null hypothesis, $B_{10} = 27.5$. Thus, it seems that many of the apparent false positives are statistical true positives that have smaller effect sizes than originally reported. The effect sizes declined by ~50% largely because the to-be-replicated findings were selected from underpowered original studies using a $p < .05$ filter, yielding inflated effect size estimates (Wilson et al., 2020). It stands to reason that many would have been detected as statistical true positives at the level of the effect size had the replication studies involved much larger *N*.

Imagine how different the impression of a replication crisis might be had the replication experiments in OSC2015 been so highly powered that the large majority of them yielded a significant effect in the same direction as the original experiments. The authors of OSC2015 reported that they had ~90% power to detect an effect size equal to the originally report effect size (see their Table 1). However, because those original effect sizes were, on average, double their true effect sizes (the results of the OSC2015 replication experiments show that to be true), the replication experiments actually had far less power than that. Indeed, a replication experiment with 90% power to detect an effect size double the actual underlying effect size would, in truth, have about 40% power, which is close to the proportion of studies that yielded a significant result. Thus, the results of OSC2015 – which perhaps more than any other findings cemented the impression of a replication crisis – are consistent with the hypothesis that all 100 of the original studies reported statistical true positives but with underlying effect sizes half of the originally reported effect sizes.

Then again, the non-significant replication effect sizes were small, on average, and our main point is that the smaller the underlying effect size, the more likely it is to reflect a theoretical false positive (even if it is a statistical true positive). Importantly, this concern applies not just to the "failed" replications but also to the "successful" replications. In fact, one of the successful replications in OSC2015 provides a case study of our main point. The largest of the original experiments selected for replication in OSC2015 was a correlational study that had a sample size of $N = 230{,}047$. Perhaps not coincidentally, the statistically significant finding in that study was associated with the smallest effect size of the OSC2015 original experiments ($r = .02$, which translates to an approximate Cohen's *d* of .05). The replication study used an even larger sample size ($N = 455{,}326$), and it obtained virtually the identical tiny effect size. Indeed, with such large *N*, the measured effect size is probably an almost exact estimate of the true underlying effect size. Because it differs from 0, it is a statistical true positive. Moreover, because both effect sizes are similar despite being small, the original effect is also validated from the "small telescopes" perspective (Simonsohn, 2015).

However, it is not validated from a theoretical perspective. If small effects tend to arise for theoretically uninteresting reasons, it follows that the smaller the non-zero underlying effect size, the less support it offers for the theory under investigation. At the level of theory, treating the non-significant OSC2015 results with an average effect size of 0.14 as false positives and this large-*N* finding with an effect size of only .05 as a true positive might be getting things backwards.

Recently, Protzko et al. (2020) reported an investigation of the replicability of 16 new discoveries from experiments that used what these authors regard as "current optimal practices," namely, high statistical power (by which they mean large *N*), preregistration, and complete methodological transparency (see also Nosek et al., 2022). The studies all used *N*s of 1,500 or more, far larger than the typical *N* used in

psychological science. Perhaps not surprisingly given such large-*N* experiments, when one lab attempted to replicate an effect discovered by another lab, the large majority replicated at $p < .05$ and with similar effect sizes. The authors concluded that "This high replication rate justifies confidence in rigor enhancing methods," such as increasing power by increasing *N*.

Such high replication rates satisfy the concerns of the statistician, whose focus is on measuring an effect size. But what about the concerns of the theoretician, whose main focus is on the truth status of a theory under investigation? Should the theoretician also take these high replication rates to mean that the original large-*N* studies reported new discoveries at the level of theory? The average effect size of the replication experiments was only $\overline{d} = 0.26$. This is slightly *smaller* than the average replication effect size of the social psychology experiments in OSC2015 ($\overline{d} = 0.33$), many of which are regarded as replication failures. One replication experiment reported by Protzko et al. (2020) had an effect size of $d \approx 0.10$, but it was clearly greater than 0 after averaging the data over three independent replications. The approach used in this study did indeed achieve a high rate of replicability, but this achievement (we argue) may come with the hidden cost of detecting theoretical false positives.

This issue appears to be coming to a head. For example, very recently, Marek et al. (2022) reported the results of brain-wide association studies (BWAS) using magnetic resonance imaging (MRI) and functional MRI (fMRI) to identify individual differences in "brain structure or function and complex cognitive or mental health phenotypes" (p. 1). Using a neuroimaging database containing results from many different labs, they were able to achieve sample sizes of thousands of individuals (vastly larger than the typical *N* of ~25 for a neuroimaging study). From our perspective, the results were entirely predictable: "BWAS associations *were smaller than previously thought*, resulting in statistically underpowered studies, inflated effect sizes and replication failures at typical sample sizes. As sample sizes grew into the thousands, replication rates began to improve and effect size inflation decreased" (p. 1, emphasis added).

To say that effect size inflation decreased is to say that the large-*N* effect sizes were small. Consider a representative example from this study: the *largest* (top 10%) region-of-interest effect sizes comparing an MRI measure of cortical thickness to a child behavioral checklist measure of psychopathology ranged from $.03 < |r| < .05$ (see their Fig. 1c, p. 3). In terms of Cohen's *d*, the range was $.06 < |d| < .10$, which is less than half of what is ordinarily considered to be a small effect size. Efforts were made to ameliorate the effects of potential nuisance factors like head motion, but what about the (inumerable) potential nuisance factors that went unnoticed? We submit that large-*N* findings like these, though highly replicable, are at significant risk of reflecting nuisance factors, not theoretically meaningful processes. We further submit that results like these portend the future of experimental psychology if it comes to value replicability above all else.

Other potential solutions to the large-*N* problem

To avoid publishing theoretical false positives masquerading as theoretical discoveries, instead of maximizing *N* to enhance replicability, we suggest trying to optimize *N*, which maximizes PPV at the level of theory. How else might the field address the problem of theoretical false positives? We consider two possibilities next.

The smallest effect size of interest The temptation to maximize *N* (to enhance replicability) while simultaneously specifying the smallest effect size of interest (to combat the insidious problem that is the main focus of this article) might seem like an attractive solution. This concept usually applies to a measured effect size (e.g., Cohen's *d*), not to an underlying effect size (Cohen's δ), but even if we apply it to δ, it does not effectively address the problem. Imagine, for example, that we specified $\delta = 0.05$ as the smallest effect size of interest. For the scenario depicted in Fig. 3, where $\lambda = 5$ and $\mu = 50$, the odds are barely greater than even that the theoretical mechanism generated the non-zero effect compared to it having been generated by a nuisance factor. As shown in Fig. 5, where $\lambda = \mu = 5$, the odds are lower still.

Critically, no matter what the smallest effect size of interest is declared to be, the larger any of the "qualifying" underlying effect sizes is, the more compelling support it provides for the theory under investigation. Therefore, in our view at least, the goal should be to conduct routine science in such a way as to select larger underlying effect sizes (in an effort to maximize PPV at the level of theory), not to minimize the damage caused by maximizing *N* (which minimizes PPV at the level of theory).

Test hypotheses that are more obviously true Another approach to maximizing the underlying effect size, thereby avoiding theoretical false positives, would be to conduct less risky science. As noted by Wilson and Wixted (2018), experimental protocols designed to test hypotheses that are already thought to be true and that are already well understood at a theoretical level (e.g., depriving people of food makes them hungry) will have larger underlying effect sizes, on average, than experimental protocols designed to test hypotheses for which there is little reason to believe they are true (e.g., people can feel the future). This is because one factor that makes an effect obvious is that its effect size

is large enough that it can be readily detected under easy-to-arrange experimental conditions and perhaps even during everyday life.

Testing already-thought-to-be-true theoretical mechanisms would result in the publication of highly replicable experiments with large underlying effect sizes. However, it would not advance our understanding at the level of theory much at all. Indeed, taking this approach too far would seem to miss the point of scientific inquiry. The goal of original science is not to confirm large effects that are already known to be true (even though it surely would achieve higher replication rates). To expand the boundaries of knowledge, it is essential to test hypotheses that are *risky* in the sense that the community of interested scientists would not deem the hypothesized result to be obvious in advance. An individual experimenter might have good reason to believe that the hypothesis is likely to be true (e.g., based on logical reasoning or a formal model that other scientists have not yet considered), but the hypothesis is risky in that interested scientists would not consider it to be obvious.

As an example, Stroop (1935) suggested that reading words is essentially automatic given extensive prior training, but naming colors is not. As Stroop put it in a dazzling display of logical reasoning: "The word stimulus has been associated with the specific response 'to read,' while the color stimulus has been associated with various responses: 'to admire,' 'to name,' 'to reach for,' 'to avoid,' etc." (p. 660). That theoretical analysis was not obviously true to interested scientists who had not thought about it to the degree that Stroop had, but it leads to a prediction: if you present the word "blue" in red letters, it will slow color-naming times (if that is the task required of participants) but not word reading times (if that is instead the task required of participants).

The predicted outcomes were observed, and the slowing effect of an incongruent color word on color naming turned out to be a large. However, despite its large size, it was not obvious to scientists in advance. The fact that the large effect becomes obvious after the fact increases our confidence in the theory that predicted it. It seems reasonable to suggest that this is how science enhances our theoretical understanding of the world, not by testing hypotheses that yield large effect sizes because they are already known to be true.

Conclusion

Many have argued before us that the null hypothesis of no difference is never strictly true (e.g., Cohen, 1990; Jones & Tukey, 2000; Meehl, 1967). With regard to quasi-experimental designs and correlational studies, that idea appears to be universally accepted. For experimental designs using random assignment, it is possible to at least imagine a methodologically perfect study in which the null hypothesis of no difference is true. However, for many studies involving random assignment, it seems likely that the experimental manipulation will have some effect on the dependent variable for theoretically uninteresting reasons. If so, according to the model of science we presented here, it would necessarily follow that the smaller the non-zero underlying effect size associated with a given experimental protocol is, the less support it offers for the theory under investigation.

Intermediate-*N* works well for original science

Our message runs counter to a compelling intuition according to which an effective solution to the replication crisis would be to run very large-*N* experiments (largely eliminating the need for NHST because almost everything would be significant) and to publish everything (thereby eliminating publication bias). At least then we would have a precise and replicable estimate of δ for every experiment. Indeed, we would, but Figs. 3 and 5 illustrate why, for theory-focused research, this approach could make a mess of the scientific literature. For original science, it would be better to have a publication mechanism that endeavors to filter out experiments associated with small δ (i.e., that endeavors to filter out theoretical false positives). In other words, as radical as it might sound, for theory-focused research, publication bias associated with the use of NHST or BNHT (using intermediate *N*) is a good thing, not a bad thing.

Götz et al. (2021) recently endorsed large-*N* studies, arguing that the resulting accumulation of small effects will provide an indispensable foundation for a cumulative psychological science. Obviously, we disagree. According to the arguments presented here, large-*N* studies will lead to the more frequent publication of $p < .05$ theoretical false positives, which will unfortunately masquerade as true positives because the effects (despite being small) will be unambiguously greater than 0 and will reliably replicate with an effect size similar to the one reported in the original study. For journals concerned with advancing theory, increasing the publication of replicable theoretical false positives (thereby decreasing PPV at the level of theory) cannot be regarded as improving science. Before going too far in that direction, the field should further debate the merits of increasing *N* without bound to maximize replicability versus optimizing *N* to maximize the probability that a claimed theoretical discovery is true.

Unlike large-*N* studies, typical-*N* studies do not have extremely high power and therefore lead to the publication of $p < .05$ findings with observed effect sizes that provide an imprecise and (on average) inflated estimate of δ. Such findings do have signal value in that their corresponding underlying effect sizes are, on average, larger

than the underlying effect sizes associated with $p > .05$ findings (Wilson et al., 2020). However, absent large-*N*, for any given $p < .05$ result, the underlying effect size remains uncertain. Many scientists are uncomfortable with that uncertainty, but the only remedy is to conduct every study with large enough *N* to precisely measure the underlying effect size. Unfortunately, that remedy might be worse than the problem it seeks to address because, if $\delta \mid H_0 > 0$, it will increase the detection of theoretical false positives and reduce PPV at the level of theory.

Large-*N* works well for replication science

Replication science is measurement focused Given our recommendation to avoid overpowered original-science experiments, are we therefore suggesting that the field must simply learn to live with the uncertainty associated with $p < .05$ findings published in the original science literature? No. Wilson et al. (2020) argued that resource-consuming large-*N* direct replications conducted by independent labs are essential (Zwaan et al., 2018). Moreover, in contrast to original science, replication science is optimized by using the largest possible *N* because it precisely estimates δ. At the replication stage, precise measurement is the only objective, and NHST and BNHT are no longer useful fictions. In other words, unlike theory-focused original science, replication science is inherently measurement-focused. Many reforms that have been proposed to address the replication crisis (the use of registered reports, abandoning NHST, conducting large-*N* experiments to measure the underlying effect size precisely, publishing everything, etc.) are much better suited to replication science than to original science.

Replicate influential experiments from original science Because using the largest possible *N* is a resource-intensive proposition, such replications are best focused on the relatively small subset of original science studies that gain currency (Wilson et al., 2020). Large-*N* direct replications are not needed for every study, most of which will have little impact. Recently, Lewandowsky and Oberauer (2020) investigated this issue by formally modeling science under a variety of replication scenarios. They concluded that replicating experiments only after they attract the community's interest "minimizes cost and maximizes efficiency of knowledge gain" (p. 1). This vision clearly differs from the competing "large-*N* publish everything" vision (Cumming, 2014), which, in our view, squanders resources while also maximizing the detection of theoretical false positives.

Although replication is an essential part of science, experimenters who replicate their own studies using large-*N* will presumably carry any nuisance factor through to the replication stage. Large-*N* replications by *independent* labs would be better because some nuisance factors that might plague the original experiment (e.g., a programming error) are unlikely to show up again. Nevertheless, the independent replication study is likely to have its own nuisance factor, which will create an effect in the same direction as the original experiment 50% of the time with sufficiently large-*N*.

Radical randomization An interesting way to address this problem would be to use "radical randomization," where many labs independently replicate a finding while varying aspects of the procedure that are not central to the theory under investigation (Baribault et al., 2018). Then again, even this approach cannot fully address the problem because some nuisance factors (e.g., an unintended effect of experimental instructions that are essential to testing the theoretical prediction) will carry through to the replication stage and be detected by others. Therefore, small underlying effects risk being theoretical false positives even when confirmed by a large-*N* study conducted by independent labs. This is why the field's focus should be on maximizing PPV at the level of theory by selecting relatively large underlying effect sizes, not embracing ever smaller effects that are at risk of being theoretical false positives.

Interpreting a precisely measured but small non-zero effect size

At the large-*N* replication stage, if it turns out that the original discovery replicates but with a small effect size (e.g., $\delta = .05$), the risk of it being a theoretical false positive is high. Then again, a theoretical true positive *can* have a small underlying effect size. Are there considerations that can help us to decide whether a small effect might be a theoretical true positive?

In agreement with sentiments expressed by Smith and Little (2018), we suggest that to the extent one can appeal to previously supported theoretical/mechanistic considerations to account for the small effect (e.g., if a well-specified quantitative theory unambiguously predicts it), the more likely it is to be a true positive at the level of theory. As explained in detail by Oberauer and Lewandowsky (2019), such models are constrained in what they predict and what they do not predict. They also rest on formalisms that are, in essence, summaries of empirical findings that have been successfully predicted in the past. As a result, the prior odds of a small underlying effect reflecting a true theoretical mechanism will be higher when the prediction is made by a formal model (one that stands on the shoulders of past research and on formal logic) compared to when it is made by a relatively imprecise and flexible verbal theory that is largely divorced from prior work (e.g., a theory of feeling the future based on quantum entanglement).

If a small underlying effect size does in fact reflect the operation of a true theoretical mechanism, then it should be possible to use that theoretical knowledge to increase its magnitude. Indeed, for theory-focused research, the goal should be to use not-yet-obvious theoretical mechanisms to generate larger underlying effect sizes (i.e., to increase statistical power that way), not to chase ever smaller ones by maximizing *N*. Despite being true positives in a purely statistical sense, the smaller the underlying effect size is, the more likely it is to be nothing more than a theoretical false positive.

Appendix

In what follows, we use $f(x)$ for the pure signal distribution (i.e., underlying effect sizes created by the theoretical mechanism of interest), $g(x)$ for the pure noise distribution (i.e., underlying effect sizes created by nuisance factors), and $h(x)$ for the signal-plus-noise distribution (i.e., underlying effect sizes created by the additive combination of the theoretical mechanism of interest and nuisance factors). The basic mathematics of adding and subtracting exponential random variables has been addressed many times previously (e.g., Bolch et al., 1998).

Signal distribution

The direction of the effect predicted by the theory is defined to be positive. When the theory being tested is true, we assume that δ_S is a random variable (x), falling in the positive range of $[0, \infty)$. For $x > 0$, we assume that $x \sim Exp(1/\lambda)$ such that the pdf of the signal distribution is:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Noise distribution

We further assume that δ_N is a random variable (x) such that $x \sim Exp(1/\mu)$. Its direction may be the same as that predicted by the theory ($x > 0$, which occurs with probability .5) or the opposite direction ($x < 0$, which occurs with probability .5). For $x > 0$, we assume that $x \sim Exp(1/\mu)$, and for $x < 0$ we assume that $x \sim Exp(-1/\mu)$. That is, the pdf of the noise distribution is:

$$g(x, \mu) = \begin{cases} .5\mu e^{-\mu x} & \text{if } x \geq 0 \text{ (same direction)} \\ .5\mu e^{\mu x} & \text{if } x \leq 0 \text{ (opposite direction)} \end{cases}$$

Signal-plus-noise distribution

The signal-plus-noise distribution is the sum of a random variable drawn from the unidirectional signal distribution and a random variable drawn from the bidirectional noise distribution. When δ_N is in the same direction as δ_S (both positive), this amounts to summing two positive exponentially distributed random variables. When δ_N is in the opposite direction as δ_S (δ_S positive, δ_N negative), this amounts to subtracting one exponentially distributed random variable from another, which is equivalent to summing them except that one of the random variables is negative.

We first consider the general case where $\lambda \neq \mu$ (and then subsequently consider the special case where $\lambda = \mu$, for which different equations apply). When the theory-based effect and the nuisance effect happen to be in the same direction (which will happen in 50% of the experiments in which the tested theoretical mechanism is true), both signal and noise are conceptualized as exponential random variables with positive values. For two positive exponentially distributed random variables, the pdf of their sum – that is, the convolution of two independent exponential distributions with rate parameters λ and μ–is given by:

$$h(x) = \frac{\lambda\mu}{\lambda - \mu}\left(e^{-\mu x} - e^{-\lambda x}\right)$$

δ_{SN} is conceptualized as a random variable (x) drawn from a distribution of this form when $\delta_N > 0$ (i.e., when the nuisance effect is the same direction as the effect predicted by the theory of interest). In other words, this equation applies to the sum of two exponential distributions in the range $(0, \infty)$. Thus, we can specify the signal-plus-noise distribution when $\delta_N > 0$ as follows:

$$h\left(x|\delta_N > 0\right) = \frac{\lambda\mu}{\lambda - \mu}\begin{cases} e^{-\mu x} - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

In the special case of $\lambda = \mu$, and when $\delta_N > 0$, the Erlang distribution applies instead:

$$h\left(x|\delta_N > 0\right) = \mu^2\begin{cases} xe^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

For the other 50% of the time in which the theory is true, the nuisance effect is in the opposite direction as the effect predicted by the theory (i.e., $\delta_N < 0$). Now, we add the signal distribution to the negative of the noise distribution, which is to say we subtract them. For two exponentially distributed random variables with rate parameters λ and μ (i.e., for the general case where $\lambda \neq \mu$), the pdf of their difference is given by:

$$h\left(x|\delta_N<0\right)=\frac{\lambda\mu}{\lambda+\mu}\begin{cases}e^{-\lambda x} & \text{if } x\geq 0\\ e^{\mu x} & \text{if } x\leq 0\end{cases}$$

In the special case of $\lambda=\mu$, this becomes:

$$h\left(x|\delta_N<0\right)=\frac{\mu}{2}\begin{cases}e^{-\mu x} & \text{if } x\geq 0\\ e^{\mu x} & \text{if } x\leq 0\end{cases}$$

To specify the overall signal-plus-noise distribution, we simply add the signal-plus-noise distribution that arises when the nuisance effect is in the same direction as the effect predicted by the theory (which occurs with probability 0.5) to the signal-plus-noise distribution that arises when the nuisance effect is in the opposite direction as the effect predicted by the theory (which also occurs with probability 0.5).

As noted earlier, for the general case where $\lambda\neq\mu$, when the effects are in the same direction (i.e., $\delta_N>0$):

$$h\left(x|\delta_N>0\right)=\frac{\lambda\mu}{\lambda-\mu}\begin{cases}e^{-\mu x}-e^{-\lambda x} & \text{if } x\geq 0\\ 0 & \text{if } x<0\end{cases}$$

And when they are in the opposite direction (i.e., $\delta_N<0$):

$$h\left(x|\delta_N<0\right)=\frac{\lambda\mu}{\lambda+\mu}\begin{cases}e^{-\lambda x} & \text{if } x\geq 0\\ e^{\mu x} & \text{if } x\leq 0\end{cases}$$

We can use these equations to specify the relevant equations in terms of the direction of the summed variable (x). If $x\geq 0$, based on the equations for $h(x|\delta_N>0)$ and $h(x|\delta_N<0)$ just above, we can write:

$$h(x|x\geq 0)=.5\left[a\left(e^{-\mu x}-e^{-\lambda x}\right)\right]+.5\left[1\times b\left(e^{-\lambda x}\right)+0\times b(e^{\mu x})\right]$$

where $a=\frac{\lambda\mu}{\lambda-\mu}$ and $b=\frac{\lambda\mu}{\lambda+\mu}$. The first bracketed term to the right of the equal sign represents the sum of the signal and noise distributions when both are in the positive direction, $h(x|x>0)$, which occurs with probability .50, and the second bracketed term on the right represents the sum of the signal and noise distributions when noise in the negative direction, $h(x|\delta_N<0)$, which occurs with probability .50. Note that one of the two terms inside those rightmost brackets is included for clarity but is multiplied by zero. It is multiplied by zero because it represents those occasions in which negative nuisance effects are larger than the corresponding signal effects, making the summed signal-plus-noise effects (x) negative. Because this expression applies to only positive x, it reduces to:

$$h(x|x\geq 0)=.5\left[a\left(e^{-\mu x}-e^{-\lambda x}\right)\right]+.5\left[b\left(e^{-\lambda x}\right)\right]$$

This is a partial pdf because it does not yet include cases where x is less than 0. For $x\leq 0$, we can analogously write:

$$h(x|x\leq 0)=.5\left[0\times b\left(e^{-\lambda x}\right)+1\times b(e^{\mu x})\right]$$

which reduces to:

$$h(x|x\leq 0)=.5b(e^{\mu x})$$

such that, over all x (i.e., from $-\infty$ to $+\infty$):

$$h(x,\lambda,\mu)=\begin{cases}.5\left[\frac{\lambda\mu}{\lambda-\mu}\left(e^{-\mu x}-e^{-\lambda x}\right)\right]+.5\left[\frac{\lambda\mu}{\lambda+\mu}\left(e^{-\lambda x}\right)\right] & \text{if } x\geq 0\\ .5\left[\frac{\lambda\mu}{\lambda+\mu}(e^{\mu x})\right] & \text{if } x\leq 0\end{cases}$$

or

$$h(x,\lambda,\mu)=\begin{cases}.5a\left(e^{-\mu x}-e^{-\lambda x}\right)+.5b\left(e^{-\lambda x}\right) & \text{if } x\geq 0\\ .5b(e^{\mu x}) & \text{if } x\leq 0\end{cases}$$

where, again, $a=\frac{\lambda\mu}{\lambda-\mu}$ and $b=\frac{\lambda\mu}{\lambda+\mu}$. This is the signal-plus-noise distribution in the general case where $\lambda\neq\mu$. It is the full pdf such that when integrated from $-\infty$ to $+\infty$, the result is 1. As an example, if $\lambda=5$ and $\mu=50$, integrating $h(x,\lambda,\mu)$ from $-\infty$ to 0 yields .045 (i.e., 4.5% of the distribution falls in the negative domain even though the theory is true), and integrating from 0 to $+\infty$ yields .955 (i.e., 95.5% of the distribution falls in the positive domain, which is the direction predicted by the theoretical mechanism of interest).

In the special case where $\lambda=\mu$, the overall signal plus noise distribution for same-direction experiments ($\delta_N>0$) is

$$h\left(x|\delta_N>0,\mu\right)=\mu^2\begin{cases}xe^{-\mu x} & \text{if } x\geq 0\\ 0 & \text{if } x<0\end{cases}$$

and the overall signal plus noise distribution for opposite-direction experiments ($\delta_N<0$) is

$$h\left(x|\delta_N<0,\mu\right)=\frac{\mu}{2}\begin{cases}e^{-\mu x} & \text{if } x\geq 0\\ e^{\mu x} & \text{if } x\leq 0\end{cases}$$

As before, we can use these equations to specify the signal-plus-noise distribution depending on the direction of the summed variable (x). If $x\geq 0$, we can write:

$$h(x|x\geq 0,\mu)=0.5\left[\mu^2(xe^{-\mu x})\right]+0.5\left[1\times\frac{\mu}{2}(e^{-\mu x})+0\times\frac{\mu}{2}(e^{\mu x})\right]$$

or, more simply,

$$h(x|x\geq 0,\mu)=0.5\left[\mu^2(xe^{-\mu x})\right]+0.5\left[\frac{\mu}{2}(e^{-\mu x})\right]$$

which further simplifies to:

$$h(x|x\geq 0,\mu)=0.5\mu e^{-\mu x}(\mu x+0.5\ )$$

For $x\leq 0$, we have:

$$h(x|x\leq 0,\mu)=.5\left[0\times\frac{\mu}{2}(e^{-\mu x})+1\times\frac{\mu}{2}(e^{\mu x})\right]$$

which simplifies to:

$$h(x|x \le 0, \mu) = 0.5\left[\frac{\mu}{2}e^{\mu x}\right] = 0.25\mu e^{\mu x}$$

Thus, for all x, we simply sum $h(x|x \ge 0)$ and $h(x|x \le 0)$:

$$h(x, \mu) = \begin{cases} .5\mu e^{-\mu x}(\mu x + 0.5\ ) & \text{if } x \ge 0 \\ .25\mu e^{\mu x} & \text{if } x \le 0 \end{cases}$$

This is the signal-plus-noise distribution in the special case where $\lambda = \mu$. This is the full pdf such that when integrated from $-\infty$ to $+\infty$, the result is 1. As an example, if $\lambda = 5$ and $\mu = 5$, integrating $h(x, \lambda, \mu)$ from $-\infty$ to 0 yields .25 (i.e., 25% of the distribution falls in the negative domain even though the theory is true), and integrating from 0 to $+\infty$ yields .75 (i.e., 75% of the distribution falls in the positive domain, which is the direction predicted by the true theory).

Likelihood ratios

We first consider the general case where $\lambda \neq \mu$. For $x \ge 0$, we noted earlier that:

$$h(x, \lambda, \mu) = .5\left[\frac{\lambda\mu}{\lambda - \mu}\left(e^{-\mu x} - e^{-\lambda x}\right)\right] + .5\left[\frac{\lambda\mu}{\lambda + \mu}\left(e^{-\lambda x}\right)\right]$$

and

$$g(x, \mu) = .5\mu e^{-\mu x}$$

Thus, for a specific positive underlying effect size, x_i, we can specify their probabilities given that "signal" is present (the theory is true):

$$P(x_i|s) = .5\left[\frac{\lambda\mu}{\lambda - \mu}\left(e^{-\mu x_i} - e^{-\lambda x_i}\right)\right] + .5\left[\frac{\lambda\mu}{\lambda + \mu}\left(e^{-\lambda x_i}\right)\right]$$

and given that only "noise" is present (the theory is false):

$$P(x_i|n) = .5\mu e^{-\mu x_i}$$

The likelihood ratio, $L(x_i)$, is given by:

$$L(x_i) = P(x_i|s)/P(x_i|n)$$

which is equal to:

$$L(x_i|x_i \ge 0) = \frac{.5a\left(e^{-\mu x_i} - e^{-\lambda x_i}\right) + .5b\left(e^{-\lambda x_i}\right)}{.5\mu e^{-\mu x_i}}$$

or:

$$L(x_i|x_i \ge 0) = \frac{a\left(e^{-\mu x_i} - e^{-\lambda x_i}\right) + b\left(e^{-\lambda x_i}\right)}{\mu e^{-\mu x_i}}$$

To find the specific value of x_i where the odds are even, we first rearrange this equation to isolate the exponential term:

$$L(x_i|x_i \ge 0) = \frac{\lambda}{\lambda - \mu} - \frac{2\lambda\mu}{\lambda^2 - \mu^2}e^{(\mu - \lambda)x_i}$$

Next, we set $L(x_i|x_i \ge 0) = 1$:

$$1 = \frac{\lambda}{\lambda - \mu} - \frac{2\lambda\mu}{\lambda^2 - \mu^2}e^{(\mu - \lambda)x_i}$$

and then solve for x_i:

$$x_i = \frac{1}{\mu - \lambda}\log\left[\frac{\lambda^2 - \mu^2}{2\lambda(\lambda - \mu)}\right]$$

which can be simplified to:

$$x_i = \frac{1}{\mu - \lambda}\left[\log\left(\frac{\lambda^2 - \mu^2}{\lambda - \mu}\right) - \log(2\lambda)\right]$$

Because $\lambda^2 - \mu^2 = (\lambda - \mu)(\lambda + \mu)$, this equation further simplifies to:

$$x_i = \frac{\log(\lambda + \mu) - \log(2\lambda)}{\mu - \lambda}$$

As an example, if $\lambda = 5$ and $\mu = 50$, the odds are even when the underlying effect size are $x_i = .0379$. If $\lambda = 5$ and $\mu = 10$, the odds are even when the underlying effect size are $x_i = .0811$.

To find the minimum odds, we start with the same equation, but instead of setting $L(x_i|x_i \ge 0) = 1$ (even odds) we set $x_i = 0$:

$$L(x_i|x_i \ge 0) = \frac{\lambda}{\lambda - \mu} - \frac{2\lambda\mu}{\lambda^2 - \mu^2}e^{(\mu - \lambda)0}$$

and then simply solve for $L(x_i|x_i \ge 0)$, which comes to:

$$L(x_i|x_i \ge 0) = \frac{\lambda}{\lambda - \mu} - \frac{2\lambda\mu}{\lambda^2 - \mu^2}$$

and reduces to:

$$L(x_i|x_i \ge 0) = \frac{\lambda}{\lambda + \mu}$$

As an example, if $\lambda = 5$ and $\mu = 50$, the minimum odds are 0.0909 (i.e., the odds are 1 to 11 that the theory is true). If $\lambda = 5$ and $\mu = 10$, the minimum odds are 0.333 (i.e., the odds are 1 to 3 that the theory is true). For $x < 0$ (i.e., when the underlying effect size is in the opposite direction predicted by the theory), the odds do not change beyond this minimum value, no matter how large the opposite-direction effect size is. The reason is that the theoretical mechanism under investigation does not contribute to

negative underlying effect sizes. Thus, different values of negative x merely reflect different magnitudes of nuisance effects (which are independent of effects caused by the theoretical mechanism of interest).

Now consider the likelihood ratio function for the special case where $\lambda = \mu$. Earlier, we noted that for $x > 0$:

$$h(x|x > 0) = 0.5\mu e^{-\mu x}(\mu x + 0.5\ )$$

and

$$g(x|x > 0) = .5\mu e^{-\mu x}$$

Thus, for a specific positive underlying effect size, x_i, we can specify their probabilities given signal (the theory is true):

$$P(x_i|s) = 0.5\mu e^{-\mu x_i}(\mu x_i + 0.5\ )$$

and given noise (the theory is false):

$$P(x_i|n) = .5\mu e^{-\mu x_i}$$

As before, the likelihood ratio, $L(x_i)$, is given by:

$$L(x_i) = P(x_i|s)/P(x_i|n)$$

which in this case is equal to:

$$L(x_i) = \left[\mu e^{-\mu x_i}(\mu x_i + 0.5)\right]/\mu e^{-\mu x_i}$$
$$L(x_i) = \mu x_i + 0.5$$

Setting $L(x_i) = 1$ and solving for x_i (i.e., finding the underlying effects size for which the odds are even) yields:

$$x_i = \frac{1}{2\mu}$$

Setting $x_i = 0$ and solving for $L(x_i)$ to find the minimum odds yields:

$$L(x_i) = 0.5$$

Positive predictive value

For a given δ and N, we are interested in $P(p < \alpha|\delta)$, which is the probability of a $p < \alpha$ outcome given δ. To compute PPV, we need to compute that value separately for when the theory is true (H_1, in which case δ was drawn from the signal-plus-noise distribution) and when it is false (H_0, in which case δ was drawn from the noise distribution). Either way, the value of interest is equal to the probability that a t-score drawn from a non-central t distribution (with degrees of freedom $\nu = N$ - 1 and non-centrality parameter $\eta = \delta\sqrt{N}$) is statistically significant. For a one-tailed t-test, $P(p < \alpha|\delta)$ is the probability that an observed t-score exceeds the positive critical criterion (t_c) for $\alpha = .05$:

$$P(p < \alpha|\delta) = \int_{t_c}^{\infty} P(t|\upsilon, \eta)\, dt$$

where $P(t|\upsilon, \eta)$ is the pdf of the Student's t-distribution. In MATLAB, the cumulative density function (cdf) for the non-central t distribution (nctcdf) can be used to compute this integral. For use with that cdf function, Equation 6 can be expressed as follows:

$$P(p < \alpha|\delta) = 1 - \int_{-\infty}^{t_c} P(t|\upsilon, \eta)\, dt = 1 - \text{nctcdf}(t_c, \nu, \eta) \quad \text{(A1)}$$

Next, we want to compute the probability of a $p < \alpha$ outcome (in the positive direction predicted by the theoretical mechanism) overall all δ for a given sample size, N, separately for noise trials and signal-plus-noise trials. The general expression is:

$$P(p < \alpha) = \int_{-\infty}^{\infty} P(x) \cdot P(p < \alpha|x)\, dx \quad \text{(A2)}$$

where x represents the underling effect size, δ, and $P(x)$ is either the noise distribution, $g(x, \mu)$, or the signal-plus-noise distribution, $h(x, \lambda, \mu)$, as defined above. For noise trials, with $\alpha = .05$, we represent the probability of a $p < .05$ outcome as P_N:

$$P_N = P(p < .05|H_0) = \int_{-\infty}^{\infty} g(x, \mu) \cdot P(p < \alpha|x)\, dx$$

and for signal-plus-noise trials, we represent the probability of a $p < .05$ outcome as P_S:

$$P_S = P(p < .05|H_1) = \int_{-\infty}^{\infty} h(x, \mu) \cdot P(p < \alpha|x)\, dx$$

From these two values, and assuming equal base rates as we have throughout, PPV for a given N is as follows:

$$\text{PPV} = \frac{P_S}{P_S + P_N}$$

We computed this value separately for N ranging from 2 to 1000 to produce the data shown in Fig. 6.

Next, we describe how we computed the mean of both the underlying and observed effect sizes associated with $p < .05$ outcomes as a function N (Fig. 7).

Expected underlying and observed effect sizes

Underlying effect size For a given sample size, N, we want the expected value of δ given a significant ($p < .05$) outcome:

$$E\left[(x|p < .05\right] = \int_{-\infty}^{\infty} xP(x|p < .05)\, dx \tag{A3}$$

where x represents the underlying effect size, δ. According to Bayes theorem:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

For our purposes, $A = x$ and $B = p < .05$. Thus:

$$P(x|p < .05) = \frac{P(x) \cdot P(p < .05|x)}{P(p < .05)}$$

It therefore follows that:

$$E\left[(x|p < .05\right] = \frac{\int_{-\infty}^{\infty} x \cdot P(x) \cdot P(p < .05|x)\, dx}{P(p < .05)}$$

or, in more complete form,

$$E\left[(x|p < .05\right] = \frac{\int_{-\infty}^{\infty} x \cdot P(x) \cdot P(p < .05|x)\, dx}{\int_{-\infty}^{\infty} P(x) \cdot P(p < .05|x)\, dx} \tag{A4}$$

$P(x)$ in the numerator and denominator of Equation A4 is now a joint function of the noise distribution, $g(x, \mu)$, and the signal-plus-noise distribution, $h(x, \lambda, \mu)$, as defined earlier. More specifically, because we assume equal base rates,

$$P(x) = 0.5\ g(x, \mu) + 0.5h(x, \lambda, \mu) \tag{A5}$$

$P(p < .05|x)$ in the numerator and denominator of Equation A4 was given earlier in Equation A1. For a given N, we computed $E[(\delta|p < .05]$ using Equation A4. Doing so separately for N ranging from 2 to 500 yielded the function relating $\bar{\delta}$ to N shown in Fig. 7.

Observed effect size A similar approach was used to compute the expected value of the observed Cohen's d, $E[(d|p < .05]$. We first computed the expected value of t for a given α level (fixed at .05) and a given N for a statistically significant outcome, $E[(t|\alpha, N, p < \alpha]$. We then divided that expected t by the square root of N to yield an expected d given a statistically significant outcome.

The relevant equations are similar to those above, but there are a few differences. Now, for example, the denominator of the expected value function is a double integral consisting of the probability of δ times the probability of drawing an observed t from the pdf of the non-central t distribution (with degrees of freedom $\nu = N$ - 1 and non-centrality parameter $\eta = \delta\sqrt{N}$) (integrated from -∞ to ∞ with respect to δ and from t_c to +∞ with respect to t). The numerator involves a similar double integral except also multiplied by the absolute value of t. That is:

$$E\left[(t|\alpha, N, p < \alpha\right] = \frac{\int_{-\infty}^{\infty}\int_{t_c}^{\infty} t\ P(x)P(y)\, dx\, dy}{\int_{-\infty}^{\infty}\int_{t_c}^{\infty} P(x)P(y)\, dx\, dy}$$

where x and y represent δ and t, respectively. $P(x)$ is given by Equation A5, and $P(y)$ is the pdf of the t-distribution (corresponding to the nctpdf function in MATLAB). For a given N, the expected value of Cohen's d is:

$$E\left[(d|\alpha, N, p < \alpha\right] = E\left[(t|\alpha, N, p < \alpha\right]/\sqrt{N}$$

We computed this value separately for each N ranging from 2 to 500, with $\alpha = .05$, yielding the function relating $\bar{d}$ to N shown in Fig. 7.

Author note We thank Stephan Lewandowsky and an anonymous reviewer for their insightful comments on an earlier version of this paper.

Declarations

Conflicts of interest We have no known conflicts of interest to disclose.

References

Asendorpt, J. B., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 2607–2612.

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568*, 435.

Bolch, G., Greiner, S., de Meer, H., Trivedi, K. S. (1998). *Queueing Networks and Markov Chains* (Chapter 1, pp. 1-34). John Wiley & Sons.

Borsboom, D., van der Maas, H. L. J., 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*, 756–766.

Button, K. S., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*, 1304–1312.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29.

Fechner, G. T. (1860). *Elements of psychophysics*. Breitkopf & Härtel.

Götz, F. M., Gosling, S. D., & Rentfrow, J. (2021). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620984483

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15–24.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*, 620–630.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411–414.

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.

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications, 11*, 358.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., … Dosenbach, N. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*. Advance online https://doi.org/10.1038/s41586-022-04492-9

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.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behavior, 3*, 221–229.

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science, 16*, 707–716.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*, 289–337.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606.

Nosek, B. A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 27.1–27.30.

Oakes, W. F. (1975). On the alleged falsity of the null hypothesis. *The Psychological Record, 25*(2), 265–272.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the Theory Crisis in Psychology. *Psychonomic Bulletin & Review, 26*, 1596–1618.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., ... & MacInnis, B. (2020). *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable.* Retrieved from https://psyarxiv.com/n2a9x/. Accessed 17 Apr 2022.

Richard, F. D., Bond Jr., C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev, **16***, 225.

Schooler, J. W. (2014). Turning the lens of science on itself: verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science, 9*, 579–584.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Smith, P. L., & Little, D. R. (2018). Small is beautiful. *Psychonomic Bulletin & Review, 25*, 2083–2101.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100–116.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology, 1*, 62.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science, 1*, 186–197.

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences, 117*, 5559–5567.

Witt, J. K. (2019). Insights into criteria for statistical significance from signal detection analysis. *Meta-Psychology, 3*, MP.2018.871.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 201–233.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioural and Brain Sciences, 41*, E120.