Supplementary Information for:

Science is not a Signal Detection Problem

Brent M. Wilson, Christine R. Harris, and John T. Wixted

Corresponding authors: Brent M. Wilson and John T. Wixted

Email: b6wilson@ucsd.edu
jwixted@ucsd.edu

This file includes:

Supplementary text
Figures S1 to S4
SI References

Table of Contents

www.pnas.org/cgi/doi/10.1073/pnas.1914237117

Increasing N

The rationale for increasing *N* is easy to understand assuming a binary world in which underlying effect sizes are categorically true ($\delta = \mu$) or categorically false ($\delta = 0$). In that case, ignoring the extra cost of running more subjects, only good things would come from increasing *N* to increase power. For example, if we represent the prior probability of the null hypothesis being true as $P(H_0)$ and the prior probability of the alternative hypothesis being true as $P(H_1)$ such that $P(H_0) + P(H_1) = 1$, the equation specifying the relationship between *PPV*, power ($1 - \beta$) and alpha (α) is:

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

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

$$PPV = \frac{R(1-\beta)}{R(1-\beta) + \alpha} \qquad (1)$$

In our simplified example above, where half the tested hypotheses are true and half are false, $P(H_0)$ equals $P(H_1)$ such that $R = 1$ (i.e., the prior odds are even). In that case, Equation 1 simplifies to:

$$PPV = [(1 - \beta)] \;/\; [(1- \beta) + \alpha] \qquad (2)$$

Expressed in words:

$$PPV = \text{power} \;/\; (\text{power} + \text{alpha})$$

From this equation, it is easy to see that as power decreases towards its minimum (i.e., as power approaches alpha), *PPV* decreases towards its minimum as well (i.e., PPV approaches 0.5 in the equal base-rate scenario). Because *PPV* decreases as power decreases, fields that typically conduct low-power experiments will have many false positives in their p < .05 literature, which is to say that *PPV* will be low. Indeed, low power is widely suspected of being a contributing factor to the 36% replication rate reported by OSC2015. If so, then increasing power by increasing N would help to correct that problem. The reason is that increasing N would increase power without affecting alpha.

Not only would *PPV* increase with higher power, but so would the average underlying effect size associated with p < .05 findings. From the binary true-vs.-false perspective, *PPV* and the average of the statistically significant effect sizes are two sides of the same coin. For example, continuing with the assumption that R = 1, if power were so low that it equals alpha (i.e., 1- β = α = .05), then half the published p < .05 results would have an underlying effect size of $\delta = \mu$ (true) and half would have an underlying effect size of $\delta = 0$ (false) such that *PPV* would equal .50, and the average effect size associated with significant findings would be $PPV \times \mu = 0.50\mu$. By contrast, if power were equal to 80%, then (according to Equation 2) .94 of the published p < .05 results would have an underlying effect size of μ and .06 would have an underlying effect size of 0, so the average underlying effect size associated with significant findings would increase to 0.94μ. The implication is clear: increasing N (thereby increasing power) would lead to a more secure scientific literature in that both more significant findings would be true and the average underlying effect size associated with those findings would be larger as well. But if underlying effect sizes are continuously distributed, ever smaller effects would be detected with increasing N.

Quantifying effect sizes

The effect sizes in OSC2015 are heterogeneous (e.g., some are based on studies that used a between-subject design, others a within-subject design) and are therefore not directly comparable to each other. Indeed, as described below, some may have been erroneously computed, making direct comparisons between the effect sizes for cognitive and social psychology (e.g., as shown in their tables) meaningless. Nevertheless, the effect-size values are distributed in *some* manner, and each effect size measure appears once as part of the original study and once again as part of a replicated study. Thus, while the effect sizes from different original experiments are not necessarily comparable to each other, the decrease from the original experiment to the replication experiment is meaningful.

The effect sizes computed by OSC2015, reported as "*r* per degree of freedom," were based on the reported test statistics from the original studies (e.g., the effect size for a comparison between two groups was computed from the reported *t*-score). This was done without regard for whether the study used a within-subject design or a between-subject design. The standard formula relating *d* to *r* is: $d = 2r/\sqrt{1-r^2}$. To instead compute *d* from a reported *t* score, the relevant formulas are $d = 2t/\sqrt{N}$ for a between-subject design, $d = t/\sqrt{N}$ for a one-sample design (as in our simulations), and $d_z = t/\sqrt{N}$ for a within-subject design, where d_z is Cohen's *d* computed from difference scores (1). When we previously converted the *r* effect sizes reported by OSC2015 to *d* effect sizes (2), we did not appreciate the fact that their *r* values had been computed the same way for both within- and between-subject designs. Thus, because cognitive psychology uses within-subject designs much more often than social psychology, this means that the cognitive effect sizes are often doubled compared to what they should be.

Consider, for example, one of the studies replicated by OSC2015, which was originally reported by Farrell (3). On page 133 of that article, a paired-samples *t*-test is reported: *t*(39) = 3.77, *p* = .001, *d* = 0.60. Using the formula $d_z = t/\sqrt{N}$ to determine the effect size from the reported *t*, we have d_z = 3.77 / √40 = 0.60, the correct value. This Cohen's *d* of 0.60 translates to an *r* of about .287. Using the incorrect formula $d_z = 2t/\sqrt{N}$ (the formula that applies to the between-subject case) to determine the effect size from the reported *t*, we have d_z = 2(3.77) / √40 = 1.20. This value translates into *r* = .516, which is the value reported by OSC2015. To us, this seems like an error. Whether or not it is an error, it means that the effect sizes for studies that used a within-subject design will be inflated relative to studies that used a between-subject design.

Even without taking into account that issue, larger underlying effect sizes observed for cognitive psychology that used within-subject designs would arise for a second reason as well (4). Recall that $\delta = \frac{u_2 - u_1}{\sigma}$, where σ represents an aggregate error term, and note that the smaller σ is, the larger the underlying effect size will be. The aggregate error term in the denominator can be conceptualized as $\sigma = \sqrt{\sigma_s^2 + \sigma_e^2}$, where σ_s^2 represents unsystematic error due to differences across subjects and σ_e^2 represents unsystematic error due to measurement error over and above individual differences. This equation applies both to a one-sample *t*-test and to an independent-sample *t*-test. However, for a within-subject design, which is commonly used in cognitive psychology, the error term needs to be expanded to include the correlation (ρ) between the underlying subject values across the two conditions: $\sigma = \sqrt{(1-\rho)\sigma_s^2 + \sigma_e^2}$. To the extent ρ is greater than 0, as it usually is in a within-subject design, it will reduce σ, thereby increasing underlying effect size (δ). Thus, the same fact that accounts for higher power in cognitive psychology (namely, a higher proportion of its studies use a within-subjects design) may also contribute to the larger underlying effect sizes observed in that field.

The upshot of all of this is that the effect sizes for cognitive and social psychology—either here or in OSC2015—are not directly comparable to each other (i.e., the difference between them is not meaningful). Indeed, still other issues complicate the comparison of effect sizes to each other even within the same field. For example, some effect sizes were based on mixed interactions, and the best method for putting effect sizes like that on a level playing field with a between-subject *t*-test is not clear. Thus, instead of trying to put all of the effect sizes on a level playing field (an impossible task without having access to the original data), we focus on the *change* in the effect size from the original to the replication study. For the same reason, we do not attach theoretical meaning to the exponential form of the distribution of underlying effect sizes used in our simulation study. It is simply the distribution that is maximally noncommittal to unknown information.

Simulating science

Specifying the relevant distributions

Because it is the maximum entropy distribution, we used the exponential as the prior distribution of underlying effect sizes. The mode of the exponential is 0, which means that for any fixed-size interval [a, a + b], where $a \geq 0$ and b is a constant, a random variable is more likely to be sampled from the interval defined by $[a = 0, b]$ than from any other interval defined by changing the value of a.[1] The implication is that many tested effect sizes—most of which will yield a non-significant result—are close to 0.

Although a principled specification of the distribution of underlying effect sizes was relatively straightforward (i.e., the exponential is the maximum entropy distribution), a principled specification of the sample-size distribution was harder to come by. The minimum value of the sample-size distribution must be 2 because, as described below, our simulated *t*-tests had $N - 1$ degrees of freedom. We could have used the geometric distribution for the sample-size distribution (the discrete analog of the exponential, with a range of 0 to ∞), adding 2 to avoid sample sizes of 0 or 1. However, in actual practice, the mode of the true sample-size distribution is unlikely to be the smallest value of 2, yet that would be the mode if we used the geometric distribution. We therefore created a more realistic sample-size distribution with a mode greater than 2 by summing two random draws from a geometric distribution and adding 2 to the result. Doing so resulted in a one-parameter sample-size distribution with the appropriate range of 2 to ∞ and a mode greater than 2, like the one shown in Fig. 3 of the main article.

[1] Although the mode is zero, all values drawn from the exponential, though possibly infinitesimal, are greater than 0. This might seem odd given that it is possible to dream up hypotheses that undoubtedly have an effect size of absolute zero. However, the effect size of interest is the hypothesis *as tested*, not the hypothesis itself. It seems reasonable to suppose that no experiment is so perfectly unbiased that the effect size would be absolute zero.

Original experiments

As illustrated in Fig. 3 of the main article, a given simulated experiment, i, involved (1) a random draw, δ_i, from an underlying exponential effect-size distribution with mean $\bar{\delta}$ and (2) a random draw from a sample-size distribution governed by a parameter g, which yielded a sample size, N_i. Those two independently sampled values (δ_i and N_i) determined the statistical power of a given simulated experiment.

To create simulated data for a given experiment, random error drawn from a unit normal distribution was independently added to δ_i for each of the N_i scores. As a concrete example, for Experiment *i,* suppose that random draws from the distributions shown in Fig. 3 yielded $\delta_i = 0.36$ and $N_i = 18$. Without measurement error, the experiment would consist of 18 scores of 0.36. In actuality, the simulated experiment involved 18 scores of 0.36 + e_{ij}, where j is the subject index and $e \sim N(0,1)$. That is, e_{ij} (the error score in Experiment i for subject j) was a random draw from a normal distribution with a mean of 0 and standard deviation of 1. A one-sample t-test was then computed from these simulated data, and the observed Cohen's d effect size was derived from that value using the formula $d_i = \frac{t_i}{\sqrt{N_i}}$. This process was repeated for a large number of simulated experiments.

In practice, instead of actually creating each individual score and computing t-score, we accomplished the equivalent by drawing a t-score from the non-central t-distribution using the MATLAB function nctrnd with parameters equal to $\delta_i\sqrt{N_i}$ (the non-centrality parameter) and $N_i - 1$ (degrees of freedom). On a very small percentage of trials, some statistically significant t_i values (and, therefore, the corresponding d_i values) ended up being inconceivably large. These values have a disproportionate effect on the estimate the expected value of d given a significant outcome. For our initial simulation, we took no steps to exclude these values so that we could accurately compute that expected value. However, it is not unreasonable to assume that, in actual practice, d_i values that are extremely large (e.g., greater than 30) would never be reported because, for example, the experimenter would assume that the result could not be accurate. Thus, we later discuss the implications of trimming the observed distribution of d_i scores, which has the effect of reducing expected regression to the mean.

The two free parameters, $\bar{\delta}$ and g, governed the two relevant distributions (Fig. 3), and they were manually adjusted separately for the cognitive and social psychology studies until the simulated $p < .05$ data approximately matched (1) the p-curves for the original experiments replicated by OSC2015 (Fig. S1) and (2) the mean of the observed Cohen's d effect size distributions for the $p < .05$ original experiments replicated by OSC2015 (Fig. S2). For cognitive psychology, the parameter settings we settled on were $\bar{\delta}_{Cog}$= 0.53 and g_{Cog} = .11, and the corresponding values for social psychology were $\bar{\delta}_{Soc}$ = 0.22 and g_{Soc} =.08.

As shown in Fig. S1, the simulation results closely approximate the p-curves for both cognitive and social psychology, and, as shown in Fig. S2, the same is true for the corresponding Cohen's d scores. The mean observed effect sizes in OSC2015 for original cognitive psychology experiments and social psychology experiments were $\bar{d}_{O(cog)} = 1.25$ and $\bar{d}_{O(soc)} = 0.77$. The corresponding values from our simulation were also $\hat{d}_{O(cog)} = 1.25$ and $\hat{d}_{O(soc)} = 0.77$.

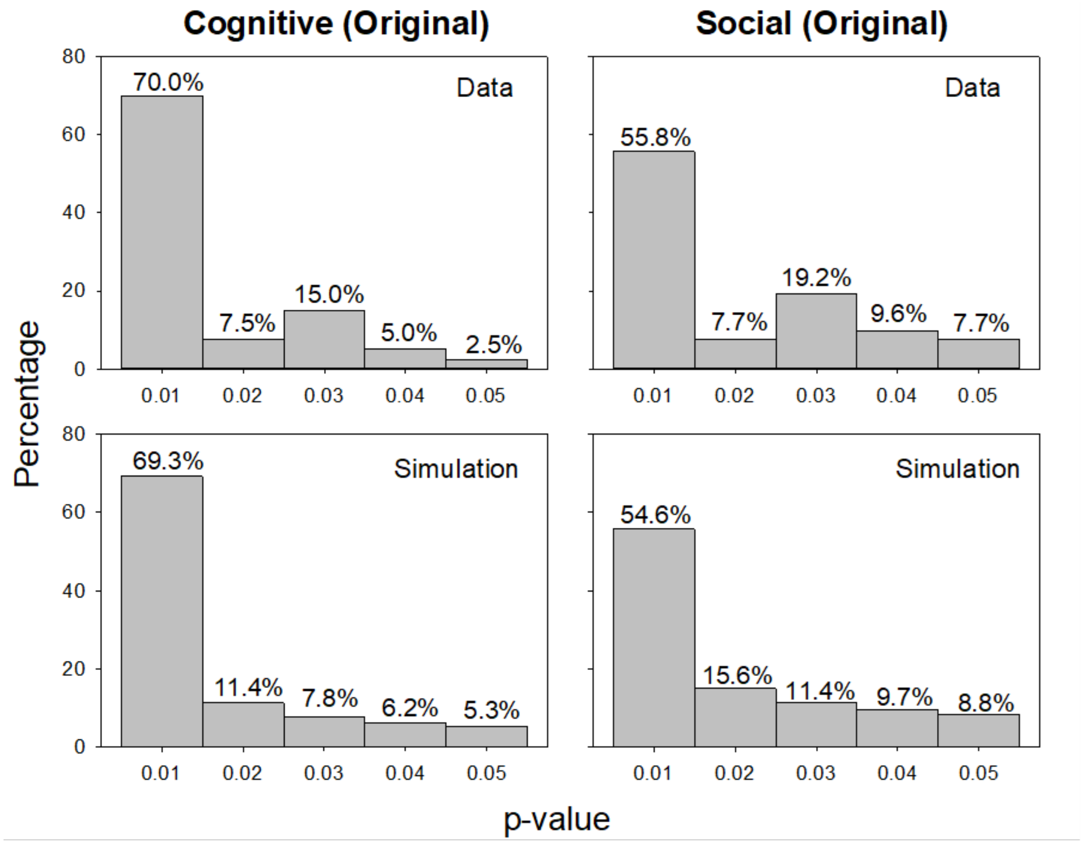

Fig S1: Observed and simulated p-curve data for original experiments from cognitive psychology and social psychology. A p-curve shows the distribution of p-values less than .05. The higher percentage of significant p-values below .01 for cognitive psychology (~70%) compared to social psychology (~55%) likely reflects the fact that power is higher for cognitive psychology, which relies on within-subject designs more so than social psychology. The empirical p-curves (upper row) show no obvious signs of QRPs such as p-hacking, though a typical-looking p-curve does not necessarily rule out other QRPs (5).

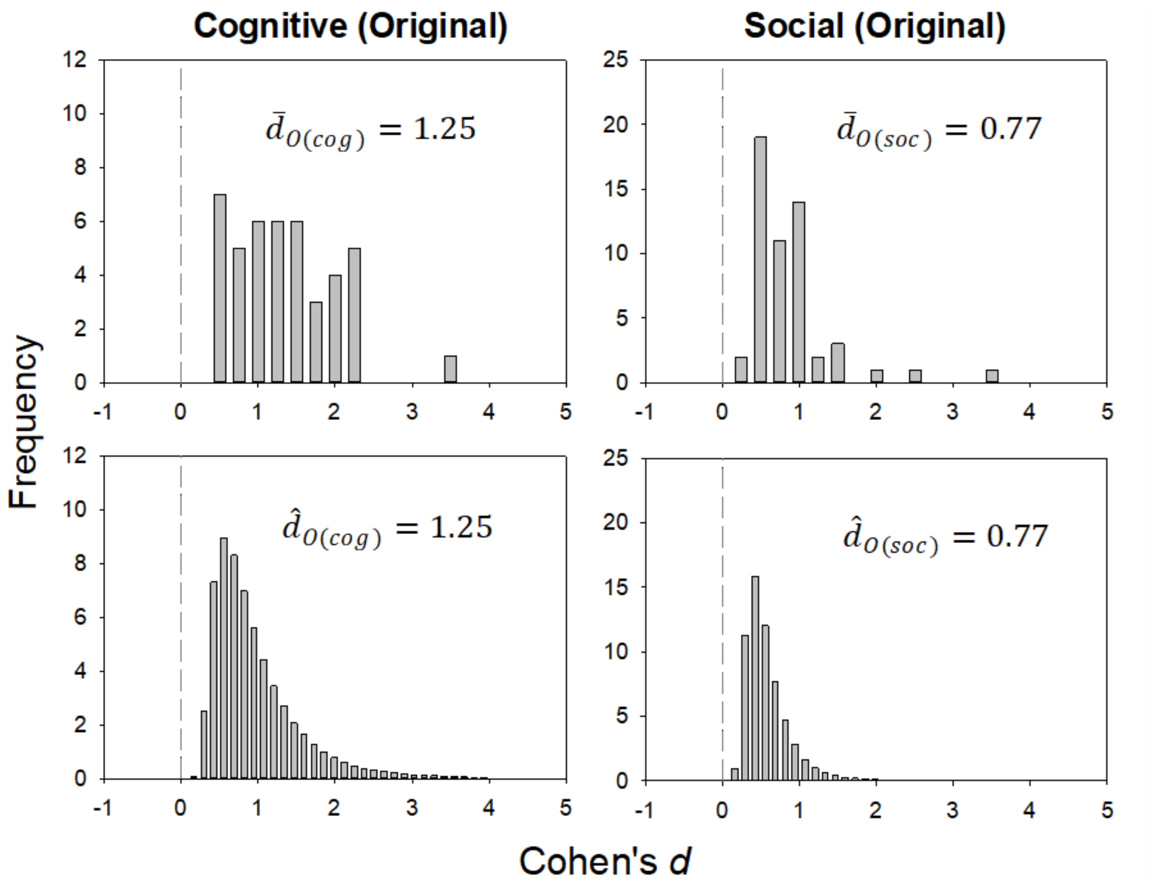

Fig S2. The top panels show the originally reported effects sizes for the cognitive psychology experiments (left) and social psychology experiments (right) from OSC (2015). The subscript "*O(cog)*" and "*O(soc)*" on the mean effect-size symbols mean "original-cognitive" and "original-social," respectively. The bottom panels show the corresponding statistically significant effect sizes from the simulation with the parameters set to $\bar{\delta}$ = .53 and *g* = .11 for cognitive psychology and $\bar{\delta}$ = .22 and *g* = .08 for social psychology. For the simulated data, the mean values reflect expected values (hence the symbol $\widehat{d}$). Note that all of the simulated effect sizes are reported here as positive even if they were, in truth, opposite in direction relative to the underlying effect size. In that case, the reported effect would be a sign error (such errors were rare in our simulated results).

Replication experiments

For the subset of simulated original experiments yielding observed effect sizes (d_i) that were statistically significant ($p < .05$, two-tailed), we performed simulated replication experiments, generating a second set of observed effect sizes. A simulated replication experiment was based on the same underlying effect size (δ_i) used for the corresponding original experiment. However, instead of drawing it from a sample-size distribution, N_i for the replication experiment was chosen to yield 90% power based on the observed effect size (d_i) of the simulated original study (following the practice used for real data in OSC2015). In the end, we had one set of (inflated) d_i values from the original studies and a corresponding set of (necessarily smaller) d_i values from the replication experiments. The results showed that about 70% of the observed decrease in effect sizes (original to replication) in OSC2015 is attributable to regression to the mean.

Next, we again simulated replication experiments but this time in an effort to capture the full decrease in observed effect sizes (original to replication). As in OSC2015, the simulated replication experiments were powered to .90 by choosing an N based on the observed Cohen's d from the original simulated experiments that achieved $p < .05$. The underlying effect size used for a given replication study (δ_{Ri}) was the underlying effect size for the corresponding original study (δ_i) multiplied by an imprecision factor (γ). The imprecision factor is intended to capture QRPs in the original studies, low-fidelity replications, or a combination of the two. With γ set to .73, the means of the simulated effect-size distributions for the replication studies ($\hat{d}_R$) are now similar to the actual means ($\bar{d}_R$). More specifically, $\hat{d}_{R(cog)} = 0.76$,

$\bar{d}_{R(cog)} = 0.75$ and $\hat{d}_{R(soc)} = 0.33$, $\bar{d}_{R(soc)} = 0.34$ (Fig. S3). In addition, in the simulated results for cognitive psychology, 52% of the simulated original $p < .05$ findings replicated at $p < .05$ (similar to the actual value of 50%), whereas for social psychology, 38% of the simulated original $p < .05$ findings replicated at $p < .05$ (somewhat higher than the actual value of 25%).

Earlier, we noted that, in our simulations of the original experiments, when the number of simulated experiments was large, a very small fraction of d_i values ended up being inconceivably large (e.g., $d_i = 500$), thereby exerting a disproportionate effect on the estimated mean observed effect size. It seems reasonable to suppose that these huge values would never end up in the scientific literature. Removing the top .0005 of the distribution largely eliminated that issue such that the maximum d_i was more in line with the maximum observed in the psychology literature (6). Taking this approach also reduced our overall estimate of regression to the mean to ~45%. Trimming observed effect sizes, while perfectly reasonable, involves the introduction of subjective exploratory assumptions that we tried to avoid as much as possible in our original simulation. In the future, efforts to quantify regression to the mean may find a principled way to deal with this issue while still managing to accurately characterize the effect-size distributions and p-curves from OSC2015.

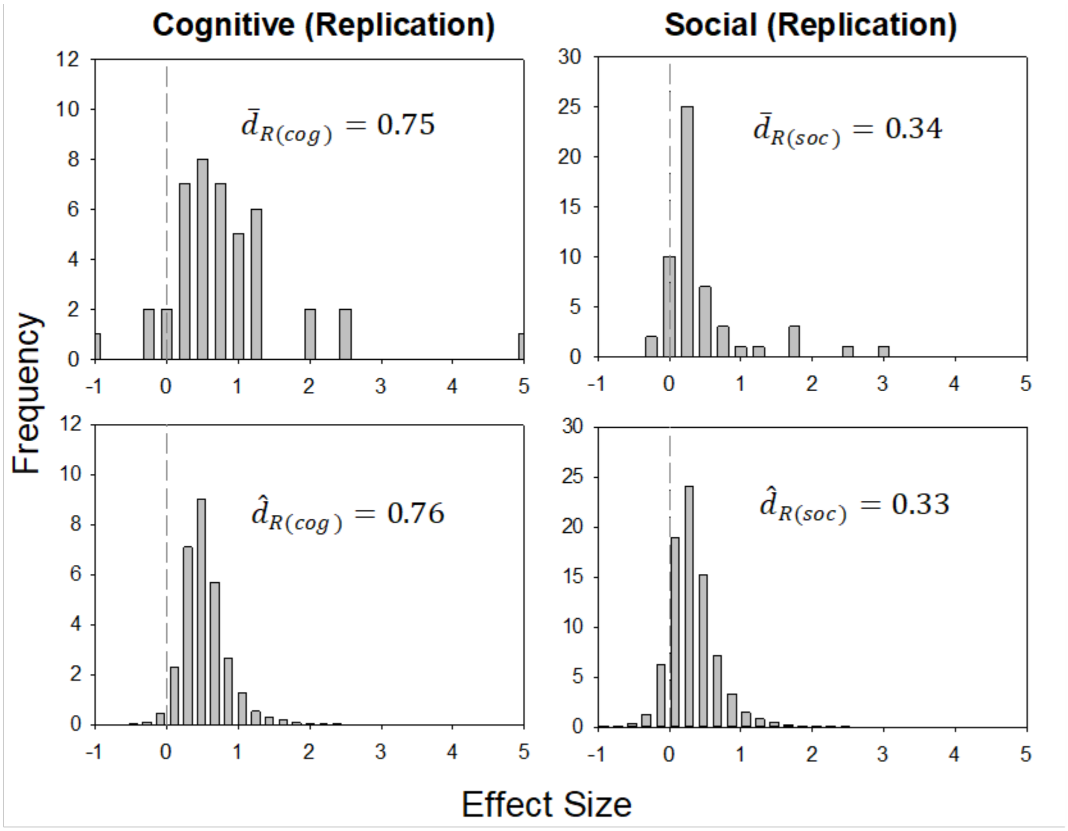

Fig S3. The two graphs in the top row show the results from the OSC (2015) replication study for the cognitive experiments (left) and social experiments (right). The two graphs in the bottom row show the corresponding results from the simulation study. Negative values mean that the corresponding original *p* < .05 effect sizes were in the wrong direction, so the replication studies would have effects in the opposite direction.

Implications for Cognitive vs. Social Psychology

In our simulations and in the OSC2015 data, significant findings from cognitive psychology were associated with larger effect sizes and were more likely to replicate compared to significant findings from social psychology. It is tempting to interpret these findings to mean that cognitive psychology is therefore a stronger science than social psychology. However, upon reflection, this is not obviously the case (2). It is certainly true that, as a general rule, large effects are more useful to both science (e.g., other scientists can readily reproduce the effect in their own research) and society (e.g., the effect can potentially have a meaningful impact on addressing a real-world problem) than small effects. However, large effects are also more likely to already be in the "encyclopedia of knowledge" than small effects. For example, the effect of depriving people of a night's sleep on how tired they are the next day would undoubtedly be large, but an experiment need not be performed to test this hypothesis because we already know it is true. Thus, a scientific discipline that focuses only on large and easily replicated effects may not be appreciably advancing knowledge despite publishing highly replicable findings (2). This why the replication rate, on its own, cannot serve as a measure of the quality of a scientific discipline. using that measure alone, the highest quality scientific discipline might the one that fails to advance knowledge at all.

Another important consideration is that OSC2015 computed effect sizes from cognitive and social psychology based on the reported test statistic in the original study (e.g., based on its reported *t* statistic). Using that approach, then, given equal effect sizes measured on a level playing field, experiments using within-subject designs would have larger *observed* effect sizes than experiments using between-subject designs. The experiments using within-subject designs would, of course, also tend to have higher power.

Almost certainly, social psychologists are more likely to investigate inherently between-subject questions (e.g., the effect of gender on cooperation and competition) than cognitive psychologists (e.g., the effect of word-frequency on recognition memory). Moreover, even when a within-subject design is technically feasible in social psychology (e.g., comparing the effect of anger vs. fear on decision-making), carryover effects often make it infeasible in practice, thereby necessitating a between-subject design. Thus, given equal resources across fields, the between-subject designs widely used in social psychology are likely to yield smaller measured effect sizes and have lower power compared to the within-subject designs more commonly used in cognitive psychology.

Because they were associated with lower power, the originally reported $p < .05$ effect sizes in social psychology were necessarily more inflated compared to cognitive psychology. Thus, when the replication studies are powered based on the originally reported effect size, as they were in OSC2015 and in our simulation study, the social psychology replications will necessarily be under-powered relative to the cognitive psychology replications. Indeed, this is precisely why, in our simulation study, cognitive experiments were more likely to replicate at $p < .05$ than social experiments. Had we powered the simulated social replications to compensate for the fact that the original Cohen's *d* scores were more inflated than the cognitive Cohen's *d* scores, the replication rates for the two fields would have been the same.

Thus, as we see it, the OSC2015 results do not have differential implications for the strength of cognitive vs. social psychology.

Optimizing *N*

When planning an original experiment that will involve NHST, power calculations designed to choose an appropriate value of *N* to ensure high power are based on faulty assumptions. As noted earlier, if underlying effect sizes are continuously distributed, then increasing power by increasing *N* is not the unambiguously good thing it would be if underlying effect sizes were categorically distributed. Indeed, if underlying effect sizes are continuously distributed, then increasing *N* too much will result in a statistically significant literature associated with underlying effects that are *less true* (i.e., closer to 0) than they are now. This seems like an important consideration because the underlying effect sizes in the current $p < .05$ literature are regarded by many to be so untrue that psychological science is in a state of a crisis.

If underlying effect sizes are continuously distributed, an alternative goal when choosing *N* might be to test the number of subjects required to *maximize* the mean of the distribution of underlying effect sizes associated with $p < .05$ findings. Critically, this goal would not be achieved by either minimizing or maximizing *N* because both of those approaches serve to reduce the mean underlying effect size associated with $p < .05$ findings. The goal of maximizing the mean underlying effect size would instead be achieved by *optimizing N*.

Assuming an exponential distribution similar to the one shown in Fig. 2C coupled with a $p < .05$ selection rule, increasing *N* would have different effects on the expected values for observed and underlying effect sizes. Fig. 5 in the main article shows the expected results assuming the mean of the underlying exponential distribution of effect sizes set to $\bar{\delta}$= .22. With regard to the underlying effect size associated with p < .05 outcomes, an inverted-U function is apparent, which means that its expected value would be maximized using an intermediate value of *N* (Fig. 5 in the main article and reproduced here in Fig. S4A). In addition, as is already well known, the average of the statistically significant reported effect sizes—that is, the average of statistically significant $|d_i|$ values—is highly inflated relative to the average δ_i when *N* is small (low power) and become less inflated as *N* increases (Fig. S4B).

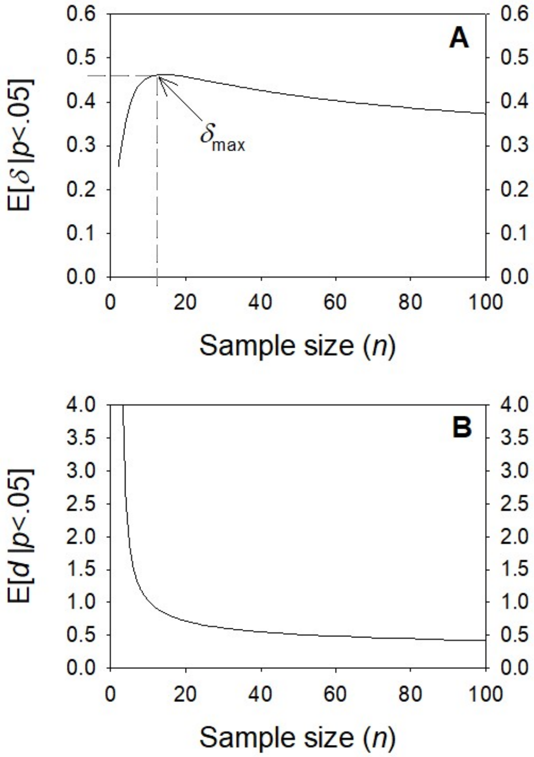

Fig S4. Expected p < .05 effect sizes with the mean of the underlying prior distribution of effect sizes set to $\overline{\delta}$ = .22 and N varied from 2 to 100. (A) Expected underlying effect size (δ) given a p < .05 outcome. (B) Expected observed effect size (d) given a p < .05 outcome.

The values depicted in Fig. S4 were computed as follows. For a given sample size, *N*, we want the expected value of *δ* given a significant (*p* < .05) outcome:

$$E[(\delta|p < .05] = \int_0^\infty \delta \cdot P(\delta|p < .05) \tag{3}$$

After computing this value for a given *N*, we can compute it for values of *N* ranging from *N* = 2 (the minimum given that *df* = *N* – 1) to some large value like *N* = 100 and then determine the *N* that yields the maximum $E[(\delta|statsig]$. The question of interest is this: which value of *N* yields the maximum expected value of *δ* given a significant (*p* < .05) outcome?

According to Bayes theorem:

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

For our purposes:

$A = \delta$

$B = p < .05$

Thus:

$$P(\delta|p < .05) = \frac{P(p < .05|\delta) \cdot P(\delta)}{P(p < .05)} \tag{4}$$

Substituting the right side of Equation 4 for $P(\delta|p < .05)$ in Equation 3, the value of interest, $E[(\delta|p < .05]$, is:

$$E[(\delta|p < .05] = \int_0^\infty \delta \cdot \frac{P(p < .05|\delta) \cdot P(\delta)}{P(p < .05)} \tag{5}$$

We used MATLAB code to compute $E[(\delta|p < .05]$ from the integral on the right side of Equation 5 (Fig. S4A). To do so, we first had to specify the numerator and denominator of Equation 4 (i.e., the rightmost term of the integral in Equation 5) more precisely, beginning with $P(p < .05|\delta)$ in the numerator. For a given *N*, $P(p < .05|\delta)$ is the probability of *p* < .05 given *δ*. It is equal to the probability that a *t*-score (*T*) drawn from a non-central *t* distribution (with degrees of freedom *ν* = *N* - 1 and non-centrality parameter $\eta = \delta\sqrt{N}$) is statistically significant. With $T{\sim}t(\upsilon, \eta)$, and for a 2-tailed *t*-test, $P(p < .05|\delta)$ is the probability that *T* exceeds either the high criterion (t_c) or falls below the low criterion ($-t_c$) under the null hypothesis for *α* = .05:

$$P(p < .05|\delta) = \int_{t_c}^{\infty} t(\upsilon, \eta) + \int_{-\infty}^{-t_c} t(\upsilon, \eta) \tag{6}$$

In MATLAB, the cumulative density function (cdf) for the non-central *t* distribution (nctcdf) can be used to compute each part of this integral. For use with that cdf function, Equation 6 can be expressed as follows:

$$P(p < .05|\delta) = \left\{1 - \int_{-\infty}^{t_c} t(\upsilon, \eta)\right\} + \int_{-\infty}^{-t_c} t(\upsilon, \eta)$$

The second term in the numerator of Equation 4, $P(\delta)$, represents the probability of drawing *δ* from the exponential prior distribution of underlying effect sizes, that is:

$$P(\delta) = (1/\tau)e^{-\tau\delta} \tag{7}$$

where $\tau = \bar{\delta}$. In MATLAB code, this is simply the probability density function for the exponential, exppdf(x,tau), where x = δ and tau = $\bar{\delta}$.

Finally, the denominator of Equation 4 is:

$$P(p < .05) = \int_0^\infty P(\delta) \cdot P(p < .05|\delta)$$

where the first term in the integral is given by Equation 7 and the second term is given by Equation 6. Again, MATLAB code was used to compute this value (i.e., the denominator of Equation 5) in the manner described above. To compute $E[(\delta|p < .05]$, the MATLAB program computed the value on the right side of Equation 5 for each value of *N* ranging from 2 to 100, with the result plotted in Fig. S4.

A similar approach was used to compute the expected value of the observed Cohen's *d*, $E[(d|p < .05]$, for each value of *N* ranging from 2 to 100 (Fig. S4B). We first computed $E[(t|p < .05]$ and 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 notable differences. Now, for example, the denominator of Equation 4 is a double integral consisting of the probability of drawing δ from the exponential prior 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 $\eta = \delta\sqrt{nN}$) times the probability that the observed *t* falls above t_c or below $-t_c$ (integrated from -∞ to +∞ with respect to *t* and from 0 to ∞ with respect to δ). The numerator involves a similar double integral except also multiplied by the absolute value of *t*. We use the absolute value of *t* on the assumption that a scientist publishing a significant finding as a new discovery would be unaware of any sign error that might exist.

Supplemental References

1 Lakens D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology,* 4:863.

2 Wilson BM, Wixted JT (2018) The prior odds of testing a true effect in cognitive and social psychology. *Adv Methods Pract Psychol Sci* 1:186-197.

3 Farrell S. (2008) Multiple roles for time in short-term memory: Evidence from serial recall of order and timing. *J Exp Psychol Learn Mem Cogn* 24*:*128–145.

4 Dunlap WP, Cortina JM, Vaslow JB, Burke MJ (1996) Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods* 1:170-177.

5 Ulrich R, Miller J (2015) p-Hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *J Exp Psychol Gen* 144: 1137–1145.

6 Szucs D, Ionnidis JPA (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol 15(3): e2000797. doi:10.1371/journal.pbio.2000797*.