Too Good to be False: Nonsignificant Results Revisited

Chris H.J. Hartgerink[1]

[1] Tilburg University, the Netherlands

WORD COUNT: 6882

Author note

This paper is version controlled and all research files are publicly available at https://osf.io/qpfnw/. Main analysis code was pre-registered.

Abstract

Significant research results in psychology have sometimes been considered “too good to be true” and possibly false positive in recent years, but we investigate whether nonsignificant results are sometimes just “too good to be false” and false negative. To this end 54,595 nonsignificant test results across eight flagship psychological journals from 1985-2013 were investigated for false negatives. We propose two ways of testing for possible false negatives, one across papers and one within a paper. All inspected journals showed evidence for false negatives, albeit to differing degrees, with 66.7% of papers reporting nonsignificant results being possibly false negative. Across the entire set of results, the false negative effect was estimated at $r \approx .2$ each year from 1985-2013. The false negative rate was estimated at 37-45%, which was also stable from 1985-2013, and the proportion of nonsignificant results reported in the psychological literature increased from 1985-2013. Sample sizes have remained similar throughout 1985-2013. This in combination with the results indicate false negatives have been far from resolved, and concern for false negatives in psychological science is warranted.

Keywords: nonsignificant, underpowered, effect size, fisher method

Too Good to be False: Nonsignificant Results Revisited

Popper's (1959/2005) falsifiability serves as one of the main demarcating criteria in the social sciences, which stipulates that a hypothesis is required to have the possibility of being proven false to be considered scientific. Within the theoretical framework of scientific hypothesis testing, accepting or rejecting a hypothesis is unequivocal, because the hypothesis is either true or false. Statistical hypothesis testing, on the other hand, is a probabilistic operationalization of scientific hypothesis testing and, in lieu of its probabilistic nature, is subject to decision errors.

Null Hypothesis Significance Testing (NHST) is the most prevalent paradigm in statistical hypothesis testing in the social sciences (American Psychological Association, 2010). In NHST the hypothesis H_0 is tested, where H_0 most often regards the absence of an effect. If deemed false, an alternative, mutually exclusive hypothesis H_1 is accepted to provide a better depiction of reality. Decisions in NHST are based on the *P*-value; the probability of the sample data, or more extreme data, given H_0 is true. If the *P*-value is smaller than the decision criterion (i.e., α; typically .05), H_0 is rejected and H_1 is accepted. Table 1 summarizes the four possible situations that can occur in NHST. The columns indicate which hypothesis is true in the population and the rows indicate what is decided based on the sample data. When the null hypothesis is true in the population and H_0 is accepted (‘H_0’), this is a true negative (upper left cell; 1-α). The true negative rate is also called specificity of the test. Conversely, when the alternative hypothesis is true in the population and H_1 is accepted (‘H_1’), this is a true positive (lower right cell). The probability of finding a positive if H_1 is true is the power (1-β), which is also called the sensitivity of the test. When H_0 is true in the population, but H_1 is accepted (‘H_1’), a Type I error is made (α); a false positive (lower left cell). Finally, when H_1 is true in the population and H_0 is accepted (‘H_0’), a Type II error is made (β); a false negative (upper right cell).

Concern about false positives has overtaken science in general and psychological science in particular. Common questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012), such as for example optional stopping, increase false positive rates considerably (Armitage, McPherson, & Rowe, 1969; Simmons, Nelson, & Simonsohn, 2011). This has increased attempts to detect false positives via replications. For example, the seminal elderly priming study (Bargh, Chen, & Burrows, 1996), where participants who were primed with senior citizens walked slower, failed to replicate (Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013), which raised doubts about whether this effect is a true positive. Nonetheless, replications show considerable variability (Klein et al., 2014), hence caution is warranted when an effect is concluded to be a false positive based on individual replications.

Previous concern about false negatives, or power, has been overshadowed by this concern for false positives, but lack of concern about false negatives is unfounded. Cohen (1962) was the first to indicate that psychological science was underpowered, which means the chance of finding an effect in the sample, if there is an effect in the population, is lower than 50%. This had barely changed thirty years later (Sedlmeier & Gigerenzer, 1989) and has not changed 50 years later (Wicherts, Bakker, & Molenaar, 2011). In other words, the concern for false negatives has been overtaken by concern for false positives, but the problem of false negatives has no evidence of being resolved. Fiedler, Kutzner, and Krueger (2012) worry that this increased focus on false positives is too shortsighted, because false negatives are more difficult to detect than false positives; they argue that negative results are less likely to be the subject of replications than positive results, decreasing the probability of detecting a false negative.

Detecting false positives and false negatives is problematic, because detection of either is afflicted by the systemic phenomena publication bias and QRPs. Even though these

phenomena and their effects on false positives have received much attention (Ioannidis, 2005; Simmons et al., 2011), the effects of these phenomena on false negatives have not, while they are plausible to have an impact there as well. Publication bias is defined as nonsignificant results having a lower probability of getting published than significant results (Greenwald, 1975), whereas QRPs biases analyses in favor of statistically significant results. These phenomena increase the difficulty to detect false positives, because publication bias decreases the probability of negative replications getting published and QRPs increase the number of (false) positive replications. As a consequence, original results are increasingly confirmed, which increases confidence in the original findings and subverts detection of false positives. For false negatives, publication bias decreases the probability of detection because the (false) negative results are less likely to be published and mostly stay under the radar to begin with. Additionally, QRPs could mask false negatives in the veil of positive results. As a consequence, more results are false positive, which, when detected, might lead to the conclusion: if false positive, then true negative. This conclusion is incorrect however, because it assumes that H_0 is true, but could easily subvert false negative detection via this route. Thus, publication bias and QRPs also increase the difficulty to detect false negatives—not just the difficulty to detect false positives. Considering that false positives and false negatives are (partly) affected by similar phenomena and that the problem of false negatives has been overshadowed, not resolved, we investigate false negatives further.

Overview of the current paper

The research question of the current paper is whether and to what degree there is evidence for false negative results in the published psychological literature. To this end, nonsignificant results from eight flagship psychological journals were inspected. First, observed effect distributions for the eight journals (combined and separately) were compared to the distribution expected if there was no effect, where a discrepancy, thus presence of false

negatives, was expected. Second, a method is proposed to test the hypothesis that H_0 is true for all nonsignificant results in a paper. Third, we inspected the power of this method, which tests whether H_0 is true in a set of results, as a function of sample size, effect size, and number of test results, in a simulation study. Fourth, we applied this method to the nonsignificant results per paper to inspect how many deviate from H_0. Fifth, the results from the method were used to tentatively estimate the false negative effect of nonsignificant results, to indicate the severity of false negatives. Sixth, we estimate the proportion of false negatives across all journals by estimating the expected amount of significant results of the method given the number of nonsignificant test results. We first review the theory to our approach of detecting false negatives, after which the method of our investigation is detailed. Subsequently, we present the results and discuss implications and limitations.

Theoretical framework

In this section we review how individual- and sets of *P*-values are distributed. We begin by explaining the function of the *P*-value and the distribution of one *P*-value, followed by the distribution of a set of *P*-values. Subsequently, we propose a way to test whether a collection of results across papers deviates from what would be expected under H_0. We also propose a way to test whether results deviate from H_0 within a paper.

Distributions of *P*-values

A single *P*-value is a function of the population effect, the observed effect and the precision of the estimate. When the population effect is zero, the distribution of one *P*-value is uniform, but when there is an effect, the distribution of one *P*-value becomes right skewed. More specifically, as the sample size increases, the precision of the estimate increases and the distribution for one *P*-value becomes increasingly more right skewed. When the observed effect increases, *P*-values also become increasingly more right skewed and an example of this is depicted in Figure 1. In other words, when there is no effect present, a *P*-value is uniform

distributed and when there is an effect present, a *P*-value is right skew distributed. As the effect or precision increases, the right skew increases. This also generalizes to aggregated *P*-values, where a set of *P*-values will be uniform distributed when there is no population effect and right skew distributed when there is a population effect, with more right skew as the population effect or precision increases. In this paper, only nonsignificant *P*-values are examined, but the tests we develop require random variables distributed in the state space [0; 1]. Hence, we apply the following transformation

$$p_i^* = \frac{p_i - \alpha}{1 - \alpha} \quad (1)$$

where p_i is the vector of untransformed *P*-values, and α is the selected significance cutoff (α = .05). This retains the distributional properties of the original *P*-values for the selected nonsignificant results.

Testing for deviation from H_0 in a set of nonsignificant results

We outline two tests to inspect whether observed, nonsignificant results deviate from H_0. First, we outline a test to inspect whether observed nonsignificant effects deviate from what would be expected under H_0, by computing an expected effect distribution. Second, we review the Fisher method, which tests whether a set of *P*-values is uniformly distributed.

To test the null hypothesis of no effect in a set of results, the observed effect distribution is compared with the expected effect distribution under H_0. Under H_0, the *P*-value distribution is known to be uniform and, given this uniformity, the probability density function of an effect can be computed via its degrees of freedom. Subsequently, the expected distribution of a set of results is the aggregate of the distributions of the individual effects given uniformity. This expected effect distribution can then be compared with the observed effect distribution to test whether there is evidence that the observed effect distribution deviates from a null effect distribution, indicating the presence of an effect.

To test for false negatives within one paper we use the Fisher method, which tests whether a set of observed *P*-values is uniformly distributed (Fisher, 1932). This technique was initially introduced as a meta-analytic technique, to synthesize results across studies and indicates whether there is evidence for deviation from H_0 in a set of results (Hedges & Olkin, 1985; Hong & Breitling, 2008). We apply this method to test whether H_0 (i.e., uniformity) holds for the nonsignificant results in a paper. In other words, the null we test with the Fisher method is that the nonsignificant results are true negative. If significant, we conclude that a false negative has occurred in that paper. The Fisher method is defined as

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \log_e(p_i) \tag{2}$$

where p_i is a vector of independent *P*-values, k is the number of values in this vector, and χ^2 has $2k$ degrees of freedom. We apply the transformed nonsignificant *P*-values (see Equation 1) in this test.

Method

Procedure

APA style test statistics were collected from 8 psychological journals. APA style is defined as the format where, in the following order, the type of test statistic is reported, the degrees of freedom (if applicable), the observed test value, and the *P*-value (e.g., $t(85) = 2.86$, $p = .005$; American Psychological Association, 2010). The inspected journals were (1) Developmental Psychology, (2) Frontiers in Psychology, (3) Journal of Applied Psychology, (4) Journal of Consulting and Clinical Psychology, (5) Journal of Experimental Psychology General, (6) Journal of Personality and Social Psychology, (7) Public Library of Science, and (8) Psychological Science. Table 2 depicts the timeframe and the number of articles downloaded per journal.

Articles from the Public Library of Science containing the subject “psychology” were downloaded with the rplos package (Chamberlain, Boettiger, & Ram, 2014) and articles from

the seven other journals were downloaded manually. We used the R package statcheck (Epskamp & Nuijten, 2013) to extract all APA reported *t*, *r*, *F*, *Z*, and χ^2 test statistics. The statcheck package not only extracts the reported test statistics, but also re-computes the accompanying *P*-value and checks for reporting errors. We used the reported *t*, *F*, and *r*-values and the re-computed *P*-values. We use the re-computed *P*-values to take into account possible rounding errors in the original reported *P*-values, which have been indicated to change statistical decisions in 15% of the results for one sample (Bakker & Wicherts, 2011).

Effect computation

The *t*, *F*, and *r*-values were all transformed into the effect size η^2, which is the explained variance for that test result, and ranges between 0 and 1. For *r*-values, this only requires taking the square (i.e., r^2). *F* and *t*-values are converted to effect sizes by

$$\eta^2 = \frac{\left(\frac{F \times df_1}{df_2}\right)}{\left(\frac{F \times df_1}{df_2}\right)+1} \tag{3}$$

where $F = t^2$ and $df_1 = 1$ for *t*-values. Adjusted effect sizes, which correct for positive bias due to sample size, were computed as

$$\eta^2_{adj} = \frac{\left(\frac{F \times df_1}{df_2}\right)-\left(\frac{df_1}{df_2}\right)}{\left(\frac{F \times df_1}{df_2}\right)+1} \tag{4}$$

which shows that when $F = 1$ the adjusted effect size is zero. For *r*-values the adjusted effect sizes were computed as (Ivarsson, Andersen, Johnson, & Lindwall, 2013)

$$\eta^2_{adj} = \eta^2 - \left(\left(1-\eta^2\right)\frac{v}{N-v-1}\right) \tag{5}$$

where *v* is the number of predictors. It is assumed that reported correlations concern simple correlations and concern only one predictor (i.e., $v = 1$). This reduces the previous formula to

$$\eta^2_{adj} = \eta^2 - \frac{1-\eta^2}{df} \quad (6)$$

where *df* is equal to *N*-2. These effects make up the observed nonsignificant effect distributions, which we compare with the expected effect distributions.

Comparing observed- and expected effect distributions for nonsignificant results

Simulations were used to approximate the expected effect distribution against which observed nonsignificant effects were tested. One simulation composed of randomly drawing a test result (with replacement) in the set of observed test results and drawing a *P*-value from a uniform distribution between 0 and 1. Based on the drawn *P*-value and the degrees of freedom, the accompanying test statistic was computed, which was subsequently used to compute the effect. Across simulations, this assumes that test results are independent. The number of simulations was pre-specified at three times the number of test results. For example, if the expected distribution is simulated for the entire set of 54,595 nonsignificant test results (see Table 2), 54,595 × 3 = 163,785 simulations were run. We compared both the entire set of results with the expected effect journal, and the results per journal. To recapitulate, the expected distribution can be computed because the *P*-value distribution is known to be uniform under H_0.

To test for differences between the expected and observed nonsignificant effect distributions, the Kolmogorov-Smirnov test was applied. This is a non-parametric goodness-of-fit test for distributions, which is based on the maximum absolute deviation between the independent distributions being compared (denoted *D*; Massey Jr., 1951). In this specific case, the fit of the observed effect size distributions was compared with the expected effect distribution for the entire set of test results and per journal. Differences in distributions between journals were not subjected to inferential significance tests, as the data are considered to be the population of *t*, *r*, and *F*-values reported in the journals.

Simulating statistical properties of the Fisher method

To examine the specificity and sensitivity of the Fisher method in testing deviation from H_0, a simulation study was conducted. The Fisher method is used to inspect whether a set of nonsignificant *P*-values within a paper are uniformly distributed (i.e., true negative). Throughout this paper, we test the Fisher method, which tests whether results are "too good to be false", with $\alpha_{fisher} = 0.10$, because tests that inspect whether results are "too good to be true" typically also use alpha levels of 10% (Francis, 2012; Ioannidis & Trikalinos, 2007). We simulate nonsignificant results from a one-sample *t*-test and outline the used procedure for one simulated Fisher method result below.

To simulate the result of the Fisher method, two simulation steps were required. First, a set of *k* nonsignificant test results ($P > \alpha = .05$) were simulated. A visual aid for simulating one nonsignificant test result is provided in Figure 2. Given sample size *N*, the one-tailed critical value of the *t*-distribution under H_0 is determined, which is subsequently used to determine β, the left tailed area under H_1 for the critical *t*-value. The H_1 distribution is determined by computing the non-centrality parameter δ as (Smithson, 2001; Steiger & Fouladi, 1997)

$$\delta = \frac{\eta^2}{1-\eta^2} N \tag{7}$$

The β value is the probability of finding a nonsignificant result (i.e., a false negative). A uniformly drawn value between 0 and β was used to compute the *t*-value under the H_1 distribution, which was then used to determine the nonsignificant, one-tailed *P*-value under the H_0 distribution. This procedure is repeated for *k* test results. Second, the set of *k* nonsignificant test results is used to compute the result of the Fisher method. Before computing the Fisher method statistic, the nonsignificant *P*-values were transformed (see Equation 1). Subsequently, the Fisher method statistic and the accompanying *P*-value were computed.

This procedure was carried out for conditions in a three-factor design, where power of the Fisher method was simulated as a function of sample size *N*, effect size η, and *k* test results. The three factor design was a 3 (sample size *N*: 33, 62, 119) by 100 (effect size η: .00, .01, .02, ..., .99) by 18 (*k* test results: 1, 2, 3, ..., 10, 15, 20, ..., 50) design, resulting in 5,400 conditions. The levels for sample size were determined based on the 25th, 50th, and 75th percentile for the degrees of freedom (df_2) in the observed dataset. Each condition contained 10,000 simulations. The power of the Fisher method for one condition was calculated as the proportion of significant Fisher method results given $\alpha_{fisher} = 0.10$, and if the power was ≥ 99.5%, power for larger effect sizes were set to 1.

Estimating false negative effects across observed nonsignificant results

To inspect how large the false negative effect is in the statistically nonsignificant test results, we used the results of the Fisher method to tentatively estimate this effect. We estimated false negative effects for the entire set of results and per journal.

The effect was estimated by simulating the expected number of significant Fisher method results given an effect size and the degrees of freedom in the observed test results. Subsequently, the expected number of significant Fisher method results was compared to the observed number of significant Fisher method results. The procedure for these simulations is the same as used in the simulation study from the previous section. However, this time only the effect size η was varied (100 levels: .00, .01, .02, ..., .99), because sample size and number of nonsignificant test results was observed. The effect size for which the expected number of significant Fisher method results showed the smallest difference with the observed number of significant Fisher method results was taken as the false negative estimate. More formally, we estimated the false negative effect across a set of results as

$$\min \frac{(O - E_\eta)^2}{E_\eta} \tag{8}$$

where η is the effect size. The resulting effect estimate is a tentative effect estimate across a set of results and should and is primarily an illustration of the degree that false negatives occur.

Estimating false negative rate for observed nonsignificant results

Considering that the results of the Fisher method are also subject to decision errors and can yield false positive indications of false negatives, we estimated the proportion of true positive indications of false negative results. To this end, we estimated the function for the expected number of significant Fisher method results across the entire dataset and applied the resulting estimates to estimate the proportion of true positive indications of false negatives compared to the number of results in total (i.e., false negative rate). We assume that the estimated function is equal across years and therefore estimate across the entire set of test results.

The expected proportion of significant Fisher method results was inspected as a function of *k* test results in two regression models. These regression models include the dichotomized significance of the Fisher method ($\alpha_{fisher} = .10$) as the dependent variable. The first model, the saturated model, was fitted with dummies for each *k* number of test results. The second model, a probit model, was iteratively estimated to find the best fitting model, where best fitting was defined as the highest explained variance. This probit model was defined as

$$1-\phi\left(\frac{z_{cv}}{\sqrt{k}},\gamma,\frac{1}{\sqrt{k}}\right) \tag{9}$$

which gives the right-tail area in a normal distribution for the value $z_{cv}/\sqrt{k}$, given mean γ and standard deviation $1/\sqrt{k}$. Here, z_{cv} is the critical value of the normal distribution given a one-tailed test with 10% alpha. The parameter γ was varied between 0 and 1.5 to obtain the best fitting probit model, where a larger γ indicates a stronger relation between the expected proportion of significant Fisher method results and the *k* number of test results. When $\gamma = 0$, there is no relation and the model is equal to the uniform null model; if $\gamma > 0$, there is a

relation between *k* and the observed power and the relation takes on an exponential shape where larger γ indicates a stronger relationship.

The expected proportion of significant Fisher method results for *k* results from the best fitting probit model and the saturated model were applied to the observed number of significant Fisher method results to estimate the false negative rate. The expected proportion of significant Fisher method results, given *k* nonsignificant test result in a paper, is denoted E_i. The false negative rate (FNR) is estimated as

$$FNR = \frac{\sum_{i=1}^{P} f_i E_i}{P} \tag{10}$$

where *P* is the number of papers, f_i the significance of the Fisher method (0 = nonsignificant; 1 = significant), and E_i the proportion expected significant Fisher method results for the *k* nonsignificant results in the *i*th article. E_i was based on either the best fitting probit model, to provide a lowerbound FNR estimate, or the saturated model, to provide an upperbound FNR estimate. In words, the estimation procedure takes the number of significant results for the Fisher method, multiplied by what would be expected given the model estimates, and divides it by the number of papers to get a proportion of false negatives.

Results

The collected dataset of *t*, *F*, and *r* values is summarized in Table 2. Figure 3 shows the distribution of observed effect sizes (in |η|) across all articles and indicates that, of the 223,082 observed effects, 43% is zero to small (i.e., $0 \leq r < .1$), 27% is small to medium (i.e., $.1 \leq r < .25$), 12% medium to large (i.e., $.25 \leq r < .4$), and 18% large or larger (i.e., $r \geq .4$). Of the full set of 223,082 test results, 54,595 (24.5%) nonsiginificant test results were selected to inspect for false negatives. Across the years, the proportion of nonsignificant results increases slowly, as depicted in Figure 4.

Compared observed- and expected effect distributions of nonsignificant results

In order to inspect whether a set of nonsignificant results across papers shows evidence for deviation from H_0, we compared the observed and expected effect distributions. If these distributions differ, we regard this as presence of false negatives. We begin by reviewing the results for the entire set of nonsignificant results, after which we inspect the sets of nonsignificant results per journal separately.

For the entire set of nonsignificant results, Figure 5 indicates that there is evidence for deviation from H_0. Even when there is no population effect, some effects will be observed due to sampling fluctuation. For example, 46% of all effects will be zero through small, as depicted by the H_0 line in the unadjusted cumulative effect distribution in Figure 5. Medium effects or smaller are expected to cover 85% of all effects if H_0 is true; large effects or smaller are expected to cover 96%. However, we observe a different distribution, where 26% of all effects are zero through small; 71% are smaller than medium, and 92% are smaller than large. Testing whether these distributions could be different due to chance, if all results were true negatives, indicates it highly unlikely ($D = 0.23$, $p < 2.2 \times 10^{-16}$). We regard this as evidence for false negatives being present in the set of nonsignificant test results.

To inspect whether evidence for false negatives is an artefact of positive bias in the effect sizes due to sample size, we also tested the entire set of adjusted nonsignificant effect sizes. The right pane in Figure 5 shows the expected and observed adjusted (nonsignificant) effect distributions. The difference between the distributions is larger and still significant ($D = 0.3$, $p < 2.2 \times 10^{-16}$), which indicates that the evidence for deviation from H_0 is not an artefact due to effect bias.

When the observed- and expected nonsignificant distributions per journal are inspected, these show similar evidence for deviation from H_0. We regard this as evidence that false negatives occur across all of these journals. There are some differences across journals, which considers the degree to which the observed distribution differs from the expected. The rank

order of the journals is as follows: Frontiers in Psychology ($D = 0.327$, $p < 2.2 \times 10^{-16}$), Journal of Experimental Psychology General ($D = 0.305$, $p < 2.2 \times 10^{-16}$), Public Library of Science ($D = 0.269$, $p < 2.2 \times 10^{-16}$), Psychological Science ($D = 0.254$, $p < 2.2 \times 10^{-16}$), Developmental Psychology ($D = 0.237$, $p < 2.2 \times 10^{-16}$), Journal of Personality and Social Psychology ($D = 0.217$, $p < 2.2 \times 10^{-16}$), Journal of Consulting and Clinical Psychology ($D = 0.179$, $p < 2.2 \times 10^{-16}$), and Journal of Applied Psychology ($D = 0.053$, $p < 7.934 \times 10^{-6}$). In other words, all journals show evidence for deviation from H_0 in the nonsignificant test results, and thus indications for false negatives.

Simulated statistical properties of the Fisher method

To inspect the power of testing for false negative results with the Fisher method, we estimated power as a function of sample size *N*, effect size η, and *k* test results. Table 3 summarizes results for the simulations of the Fisher method for small- and medium effect sizes. Results for all 5,400 conditions can be found on the OSF (osf.io/qpfnw); we try to summarize them here with reference to our hypotheses.

Before summarizing the results depicted in Table 3, we explain what the number in one cell means. For example, take the second cell in the third row, which depicts a value of .2667. This indicates that when there are two nonsignificant test results, both from samples with 62 respondents and the population effect is $r = .1$, the Fisher method will correctly identify false negatives in 26.67% of the cases. In other words, the higher the value in the cell, the more sensitive the Fisher method is in detecting false negatives. This value was computed for a variety of population effects, sample sizes, and number of nonsignificant results. As a comparison value, 1-β depicts the power of a correlation test given the same sample size and population effect.

We hypothesized that the power of the Fisher method increases as sample size, effect size, or the number of test results increase and the results confirm these hypotheses. Table 3

shows the power for small- and medium population effects, but does not show results for zero- and large population effect sizes, because results were highly similar for these effect sizes. For zero population effects, results confirmed that the Fisher method retains alpha sensitivity, given a 90% confidence interval around the 10% alpha level (i.e., [.095; .105]; Agresti & Coull, 1998). For large population effects, power was at least 97% and therefore omitted. For small- and medium population effects, power of the Fisher method was more variable and therefore included. More specifically, power of the Fisher method rapidly approaches high power when there is a medium population effect, such as 77% power when 2 results from small samples (N = 33) are inspected. For small population effects however, 50 test results based on small samples give only 69% power. However, for small to medium effects (r = .17; N = 33), eight results provide 76% power. These results indicate that the Fisher method can detect false negatives with rapidly increasing power as effect size, number of results, or sample size increases—however, it should not be expected to work wonders when inspecting small effects in small samples with only few test results.

False negatives detected with the Fisher method

The Fisher method was applied to all nonsignificant test results per paper, to inspect whether these results deviate from H_0. Table 4 shows the amount of significant Fisher method results per journal and per k number of nonsignificant test results and the first row indicates the number of papers that report no nonsignificant results. Overall results indicate that for 6,951 papers nonsignificant results deviate from H_0, indicating false negatives. This is 47.1% of all papers, and 66.7% of papers that report at least one nonsignificant results.

Inspecting these results per journal shows that journals differ in the amount of significant Fisher method results, which indicate that the nonsignificant results deviate from H_0. The minimum proportion of significant Fisher method results is 49.4% (Journal of Applied Psychology), for papers that report at least one nonsignificant result. This indicates

that at least 49.4% of all papers reporting nonsignificant results are not in line with H_0, and these results are false negative. All other journals show higher proportions of papers that deviate from H_0, with Journal of Personality and Social Psychology on top (81.3%). Researchers should thus be wary to interpret negative results in journal articles as a sign that there is no effect; it is likely to be a possible false negative.

Estimated false negative effects across observed nonsignificant results

In order to investigate the degree to which false negatives occur, false negative effects were estimated. To this end, the simulation procedure used in the simulation study was applied to the observed test results. Based on this procedure, outlined in the method section, the estimates varied between correlations of .11 and .32. This corresponds with explained variances of 1.2% and 10.2%; small and medium-large effects. Across all journals, the false negative effect was estimated at $r = .17$. The journal rank order for the estimated false negative effects is as follows, when sorting from largest to smallest estimate: Public Library of Science ($r = .32$), Frontiers in Psychology ($r = .20$), Psychological Science ($r = .19$), Developmental Psychology ($r = .17$), Journal of Personality and Social Psychology ($r = .16$), Journal of Applied Psychology ($r = .15$), Journal of Experimental Psychology General ($r = .12$), and Journal of Clinical and Consulting Psychology ($r = .11$). These estimates are tentative and illustrate the size of the false negative effect that is being neglected. Nonetheless, the estimated false negative effects were all larger than zero, indicating false negative effects of different degrees across journals.

To inspect whether the estimated false negative effect changes over the years, the estimation procedure was carried out per year for the entire set of nonsignificant results. The upper panel in Figure 7 indicates that the estimated false negative effect is stable across 1985-2013 at $r \approx .2$. This indicates that the general trend for the degree to which false negatives occur (but not necessarily the rate) is stable over time. This could be caused by lack of

changes in the sample size over time (lower panel), which confirms earlier findings of stable sample sizes used in the psychological sciences (Marszalek, Barber, Kohlhart, & Holmes, 2011).

Estimated false negative rate for observed results

To estimate the proportion of true positive indications of false negatives in the observed dataset, the expected number of significant Fisher method results was estimated as a function of *k* test results. Results from the probit model indicate there is a direct relation between the number of results and the expected frequency of significant Fisher method results. This is additional evidence for presence of false negatives. More importantly, the optimal regression solution was found for the probit model with $\gamma = .7$, in which the model explained 2.6% of the observed variance in significance of the Fisher method. This is more than the null model (0%), but less than the saturated model (67.7%).

Based on these regression models, the proportion of true positive indications for false negatives was estimated. Across all observed test results, this resulted in a lowerbound estimate of 2,589 and an upperbound estimate of 3,136 true positive Fisher method results, of the 6,951 significant Fisher results. In other words, 2,589-3,136 papers are estimated to show truly false negative results. This is equal to an overall false negative rate of 37-45% of all papers that report nonsignificant results. Estimated false negative rates per year indicate that these are relatively stable across the years (Figure 8). For the most recent year, 2013, the false negative rate is estimated at 33-38%. These results indicate that the estimated false negative rate, when taken across all years, is larger than the base rate of 20% by a factor of 1.85-2.25 and corroborates the stability of the false negative effect.

Discussion

The current paper investigated whether nonsignificant results in eight major psychological science journals deviate from what would be expected under H_0, to inspect for

false negatives. Under H_0, the uniform *P*-value distribution was used to develop two test procedures to test for false negatives. To inspect whether nonsignificant results from a set of papers show indication for false negatives, a procedure to compute the expected H_0 effect distribution was proposed. To inspect whether nonsignificant results from one paper indicate a false negative, the Fisher method was proposed. The sensitivity of the Fisher method to detect false negatives was investigated, and the false negative effect and the proportion of false negatives were estimated.

The results of this investigation show that there is strong evidence for false negatives across all eight major journals in psychological science. The Fisher method proved a sensitive test to inspect for false negatives in our simulation study (see Table 3). Applying the Fisher method to the observed results from the eight journals indicated that 66.7% of all papers that report nonsignificant results are possibly false negative (Table 4). Journals differed in the proportion of possibly false negative results, where proportions were at least ~50% and at most ~80%. Estimates of false negative effects and false negative rates indicated that this has been highly similar across the last 30 years, which indicates that the problem of false negatives has not been resolved. We estimated the true false negative rate across the entire set of results at 37-45%. Fiedler et al. (2012) provided thorough argumentation for increased attention to false negatives and this paper provide evidence that this call is warranted.

Limitations

Before discussing further implications of these findings, some possible caveats need to be addressed. These are related to the way statcheck (Epskamp & Nuijten, 2013) extracts test statistics, whether possible dependency of *P*-values poses a threat to the validity of results, and the estimation procedure for the false negative rate.

First, statcheck extracts inline, APA style reported test statistics, but does not include results included from tables or results that are not reported as the APA prescribes. Even

though factorial experiments are common in several fields of psychological science, which are most often represented with inline test results, some fields use regression models that are more aptly summarized in table form. Such tables are not included in the results extracted by statcheck, because they do not provide the necessary information to recompute *P*-values. For example, a regression table might provide a *t*-value for the included predictors, but no degrees of freedom and only an asterisk for significance. Considering that statcheck's main aim is to recompute and check reported *P*-values, it makes sense that such incomplete data are not included.

Second, statcheck extracts test statistics from an article but does not inspect the origins of the test statistics (e.g., different studies within a paper). As a result, *P*-values extracted from one paper might be from one study, resulting in dependency between test results. Bland (2013) has illustrated that uniformity does not hold for H_0 when *P*-values are dependent. Based on the intraclass correlation (ICC), we can inspect how much dependency is present for the *P*-values within a paper. If ICC = 1, this indicates full dependency, whereas ICC = 0 indicates full independence. For the set of observed results, the ICC for *P*-values was .001, indicating near full independence of *P*-values within a paper. In other words, the assumption of independent test results was not violated and is of no further concern.

Lastly, the models used to estimate the expected frequency of significant Fisher method results, and subsequently the proportion of true indications for false negatives, were assumed to be equal across all years. It is possible that the stable false negative rate is partly due to this assumption. However, in hindsight, results (false negative effect, sample size and false negative rate) corroborate each other in their stability, affirming the result that false negatives occur approximately just as frequently in recent years as they did in the previous decades. Hence, this concern is ameliorated by corroboration of the trend.

Implications

In recent years false positive errors have received much attention, whereas only a few papers have paid attention to the importance of false negatives explicitly (e.g., Fiedler et al., 2012; Lakens & Evers, 2014). Implicitly, the importance put on power in replications (Brandt et al., 2013) was stressing the importance of controlling false negatives in replications and even though these implicit elements have trickled down into some policy changes (e.g., Eich, 2013; Journal of Experimental Social Psychology, 2014), active discussion of false negative rates beyond the scope of replications has lagged behind on the discussion of false positives. This largely neglects previous concern about power (Bakker, Van Dijk, & Wicherts, 2012; Cohen, 1962; Marszalek, Barber, Kohlhart, & Holmes, 2011; Sedlmeier & Gigerenzer, 1989), which was even addressed by an APA Statistical Task Force at the time (Wilkinson & APA Task Force on Statistical Inference, 1999). Their recommendation of increasing power bore no fruit (Marszalek et al., 2011) and our results confirm that the original concern about power/false negatives is unresolved. Moreover, the results indicate that there is no reason to even believe the problem has been partially solved. Both the estimates for the rate- and the degree of false negatives have remained stable from 1985-2013. In other words, there is no reason to believe the problem of false negatives has been resolved, but decreased debate is reason to believe a problem is being neglected.

However, what has changed is the amount of negative results being reported. Our data show that more nonsignificant results are reported throughout the years (see Figure 4). This is contrary to other findings that indicate that relatively more significant results are being reported (Fanelli, 2011; Sterling, Rosenbaum, & Weinkam, 1995; Sterling, 1959). Even though we note that other research focused on main results and we focus on results in general, the relative number of negative results is increasing and hence the relative number of false negatives is increasing. If anything, this should increase concern about false negatives, because more false negative results are being published and valuable hypotheses are possibly

disregarded because of it (Fiedler et al., 2012). Nonetheless, the increase in published negative results is positive, as it seems the field is not shying away from publishing negative results per se, as so many have proposed before (e.g., Greenwald, 1975; Nosek, Spies, & Motyl, 2012; Rosenthal, 1979; Schimmack, 2012).

After the commotion caused by the special issue on replications in the journal *Social Psychology*, false negatives have received some minor attention on social media and blogs. Wilson (2014) was concerned about failed replications being false negative. More specifically, he was concerned that the replication researchers biased the results towards nonsignificance (i.e., *Q*-hacking), which is a reasonable concern. However, the concern of *Q*-hacking extends to non-replication studies when non-effects are hypothesized (e.g., 'we expect no effect of gender on Y'). Our results also indicate, as opposed to Wilson's (2014) viewpoint, that false negatives are also a problem for eight major journals in the psychological sciences, which show evidence for false negatives. Additionally, Mitchell (2014) proposed that negative results have no philosophical value, which rejects any concern about false negatives. However, Mitchell's (2014) viewpoint is a theoretical one that neglects the probabilistic basis of results in the NHST paradigm. However valid his points are in theory, in NHST (false) negative results are informative because they can be negative by chance alone.

Conclusion

Taken together, we regard these results as considerable proof for presence of- and concern for false negatives across the entire psychological sciences. The overall false negative rate was estimated at 37-45% and we hope this numeric estimate instigates discussion of false negatives by making the problem more effable. Moreover, researchers can easily apply the Fisher method to test for false negative effects in their own research, which could prove especially helpful in a set of mixed results of an effect (e.g., two nonsignificant- and one significant main effect). Further research could expand the scope of the investigation into

false negatives by including additional journals, refining our estimation procedures, and developing a Fisher method that is robust for dependency in the *P*-values. To facilitate this, our data and procedures are provided on the OSF page for this paper (osf.io/qpfnw). We conclude by saying that discussing false negatives is not zero-sum to discussing false positives, and consider the current results a firm encouragement to add false negatives to the current discussion on error-control.

References

Agresti, A., & Coull, B. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician, 52*, 119–126.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, D.C.: American Psychological Association.

Armitage, P., McPherson, C., & Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, 132*, 235–244. Retrieved from http://www.jstor.org/stable/10.2307/2343787

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

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–78. doi:10.3758/s13428-011-0089-5

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

Bland, M. (2013). Do baseline P-values follow a uniform distribution in randomised trials? *PLoS ONE*, *8*, e76010. doi:10.1371/journal.pone.0076010

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van ’t Veer, A. (2013). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi:10.1016/j.jesp.2013.10.005

Chamberlain, S., Boettiger, C., & Ram, K. (2014). rplos: Interface to PLoS Journals search API. Retrieved from https://github.com/ropensci/rplos

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS One*, *7*, e29081. doi:10.1371/journal.pone.0029081

Eich, E. (2013). Business not as usual. *Psychological Science*, *25*, 3–6. doi:10.1177/0956797613512465

Epskamp, S., & Nuijten, M. B. (2013). statcheck: Extract statistics from articles and recompute p values.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. doi:10.1007/s11192-011-0494-7

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669. doi:10.1177/1745691612462587

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (10th ed.). Edinburgh, United Kingdom: Oliver and Boyd.

Francis, G. (2012). Too good to be true : Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. doi:10.3758/s13423-012-0227-9

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PloS One*, *8*, e72467. doi:10.1371/journal.pone.0072467

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. London, United Kingdom: Academic Press.

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, *24*, 374–82. doi:10.1093/bioinformatics/btm620

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P., & Trikalinos, T. a. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253. doi:10.1177/1740774507079441

Ivarsson, A., Andersen, M. B., Johnson, U., & Lindwall, M. (2013). To adjust or not adjust: Nonparametric effect sizes, confidence intervals, and real-world meaning. *Psychology of Sport and Exercise*, *14*(1), 97–102. doi:10.1016/j.psychsport.2012.07.007

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–32. doi:10.1177/0956797611430953

Journal of Experimental Social Psychology. (2014). JESP Editorial Guidelines. Retrieved from http://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-editorial-guidelines/

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. a. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278–292. doi:10.1177/1745691614528520

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, *112*, 331–348. doi:10.2466/03.11.pms.112.2.331-348

Massey Jr., F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*, 68–78.

Mitchell, J. (2014). On the emptiness of failed replications. Retrieved from http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

Nosek, B. a., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. doi:10.1177/1745691612459058

Popper, K. (1959). *The logic of scientific discovery*. London, United Kingdom: Routledge.

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

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–66. doi:10.1037/a0029487

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–66. doi:10.1177/0956797611417632

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632. doi:10.1177/00131640121971392

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108–112.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, *6*(11), e26828. doi:10.1371/journal.pone.0026828

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wilson, T. (2014). Is There a Crisis of False Negatives in Psychology? Retrieved from https://timwilsonredirect.wordpress.com/2014/06/15/is-there-a-crisis-of-false-negatives-in-psychology/

Table 1

Summary table of possible NHST results.

		Population	
		H_0	H_1
Decision	'H_0'	1-α *True negative*	β *False negative* *[Type II error]*
	'H_1'	α *False positive* *[Type I error]*	1-β *True positive*

Note. Columns indicate the true situation in the population, rows indicate the statistical conclusion based on sample data. The true positive probability is called power and sensitivity, whereas the true negative rate is also called specificity.

Table 2

Summary table of articles downloaded per journal.

Journal (Acronym)	Timeframe	Articles	Results	Mean results/article	Significant	% Significant	Nonsignificant	% Nonsignificant
Developmental Psychology (DP)	1985-2013	2,782	30,920	13.5	24,584	79.5%	6,336	20.5%
Frontiers in Psychology (FP)	2010-2013	3,519	9,172	14.9	6,595	71.9%	2,577	28.1%
Journal of Applied Psychology (JAP)	1985-2013	3,381	11,240	9.1	8,455	75.2%	2,785	24.8%
Journal of Consulting and Clinical Psychology (JCCP)	1985-2013	1,184	20,083	9.8	15,672	78.0%	4,411	22.0%
Journal of Experimental Psychology: General (JEPG)	1985-2013	5,108	17,283	22.4	12,706	73.5%	4,577	26.5%
Journal of Personality and Social Psychology (JPSP)	1985-2013	2,307	91,791	22.5	69,836	76.1%	21,955	23.9%
Public Library of Science (PLOS)	2003-2013	2,126	28,561	13.2	19,696	69.0%	8,865	31.0%
Psychological Science (PS)	2003-2013	10,303	14,032	9.0	10,943	78.0%	3,089	22.0%
	Totals	*30,710*	*223,082*	*14.3*	*168,487*	*75.5%*	*54,595*	*24.5%*

Note. Significance level = .05, two-tailed

Table 3

Summary table of Fisher method power simulations for small- and medium effects, for different sample sizes and number of test results. Each condition consists of 10,000 simulations.

	r = .1			*r* = .25		
	N = 33	*N* = 62	*N* = 119	*N* = 33	*N* = 62	*N* = 119
1-β	0.2319	0.3046	0.4202	0.5466	0.7517	0.9291
k = 1	0.1512	0.2110	0.3410	0.5752	0.8516	0.9833
k = 2	0.1746	0.2667	0.4591	0.7793	0.9778	1
k = 3	0.2008	0.3167	0.5717	0.8935	1	1
k = 4	0.2077	0.352	0.6587	0.9482	1	1
k = 5	0.2287	0.3897	0.7194	0.9748	1	1
k = 6	0.2510	0.4336	0.7842	0.9899	1	1
k = 7	0.2585	0.4710	0.8336	0.9953	1	1
k = 8	0.2801	0.5136	0.8709	0.9979	1	1
k = 9	0.2984	0.5298	0.8945	1	1	1
k = 10	0.3035	0.5702	0.9178	1	1	1
k =15	0.3624	0.6912	0.9798	1	1	1
k = 20	0.4291	0.7804	0.9958	1	1	1
k = 25	0.4898	0.852	0.9995	1	1	1
k = 30	0.5308	0.8936	1	1	1	1
k = 35	0.5780	0.9303	1	1	1	1
k = 40	0.6214	0.9530	1	1	1	1
k = 45	0.6538	0.9661	1	1	1	1
k = 50	0.6855	0.9762	1	1	1	1

Note. Power values of 1 are rounded power values when > .995. 1-β is the power for a correlation test given sample size and effect size, one-tailed, α = .10.

Table 4

Summary table of Fisher method results across journals. A significant Fisher method result is indicative of a false negative.

		Overall	DP	FP	JAP	JCCP	JEPG	JPSP	PLOS	PS
	Nr. of papers	14,759	2,283	613	1,239	2,039	772	4087	2,164	1,562
k = 0	Count	4,340	758	133	488	907	122	840	565	527
	%	29.4%	33.2%	21.7%	39.4%	44.5%	15.8%	20.6%	26.1%	33.7%
k = 1	Significant	57.7%	66.1%	41.2%	48.7%	58.7%	51.4%	66.0%	47.2%	56.4%
	Count	2,510	433	102	238	380	109	556	339	353
k = 2	Significant	60.6%	66.9%	50.0%	36.3%	57.7%	66.7%	75.2%	51.6%	57.1%
	Count	1768	293	64	157	227	81	424	289	233
k = 3	Significant	65.3%	69.8%	57.6%	53.1%	54.4%	77.1%	80.6%	47.8%	60.2%
	Count	1257	199	66	98	125	83	341	184	161
k = 4	Significant	68.7%	75.0%	63.8%	53.1%	69.7%	67.9%	81.4%	52.7%	62.5%
	Count	892	128	47	64	89	56	264	148	96
$5 \leq k < 10$	Significant	72.3%	71.2%	67.7%	56.7%	66.3%	71.2%	87.1%	52.4%	63.0%
	Count	2,394	326	124	134	208	163	898	368	173
$10 \leq k < 20$	Significant	77.7%	76.9%	67.7%	60.0%	72.4%	81.2%	88.1%	57.3%	81.0%
	Count	1,280	121	65	55	87	117	596	218	21
$k \geq 20$	Significant	84.0%	76.0%	53.8%	60.0%	87.5%	80.5%	94.0%	69.1%	0.0%
	Count	324	25	13	5	16	41	168	55	1
All	Significant	47.1%	46.5%	45.1%	29.9%	34.3%	59.1%	64.6%	38.4%	39.3%
	Significant $k \geq 1$	66.7%	69.6%	57.6%	49.4%	61.7%	70.2%	81.3%	51.9%	59.2%
	Count	6,951	1,061	277	371	699	456	2,641	831	615

Note. DP = Developmental Psychology; FP = Frontiers in Psychology; JAP = Journal of Applied Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEPG = Journal of Experimental Psychology: General; JPSP = Journal of Personality and Social Psychology; PLOS = Public Library of Science; PS = Psychological Science.

Figure 1

Distribution of a *P*-value as a function of effect size given sample size $N = 100$.

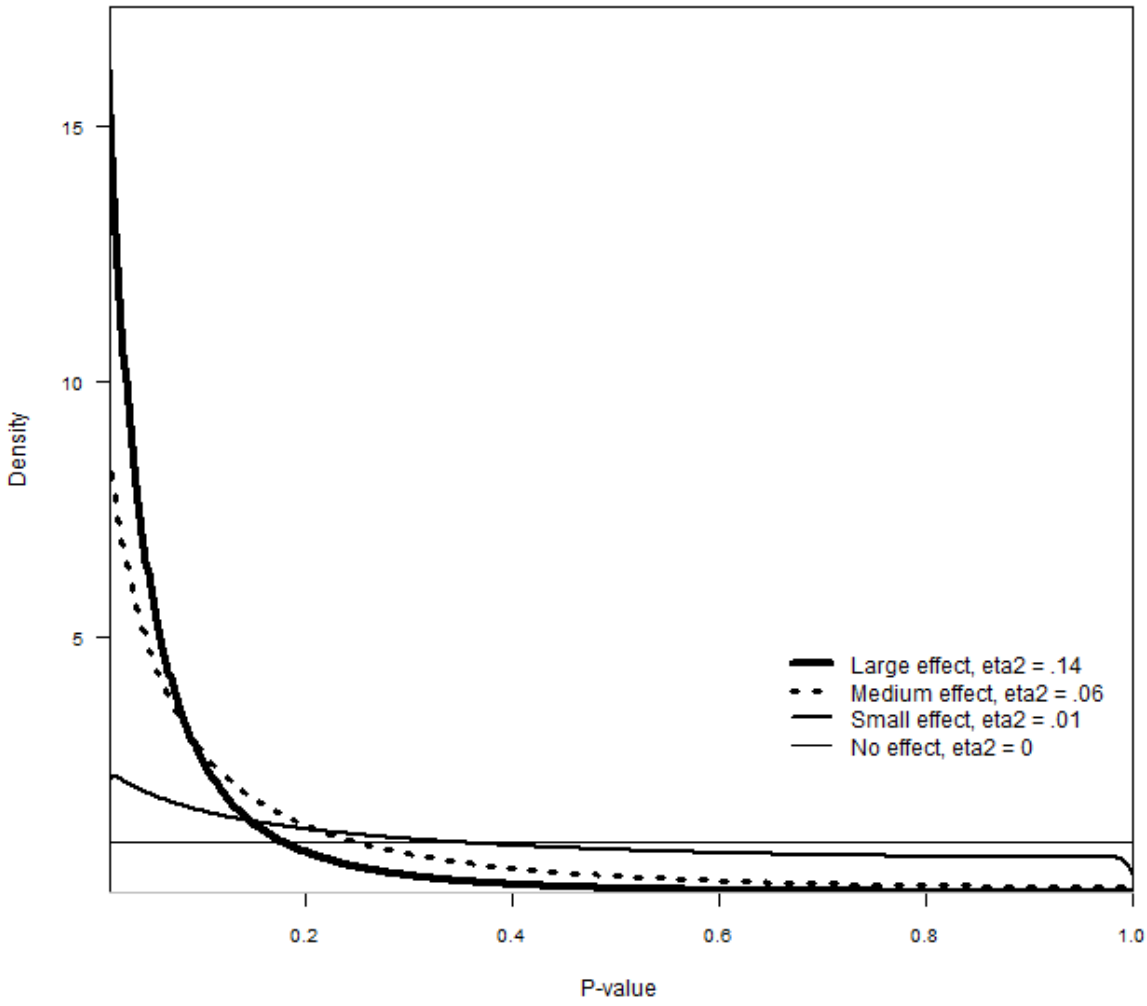


Figure 2

Visual aid for for simulating one nonsignificant test result. The critical value from H_0 (left distribution) was used to determine β under H_1 (right distribution). A value between 0 and β was drawn, *t*-value computed and *P*-value under H_0 determined.

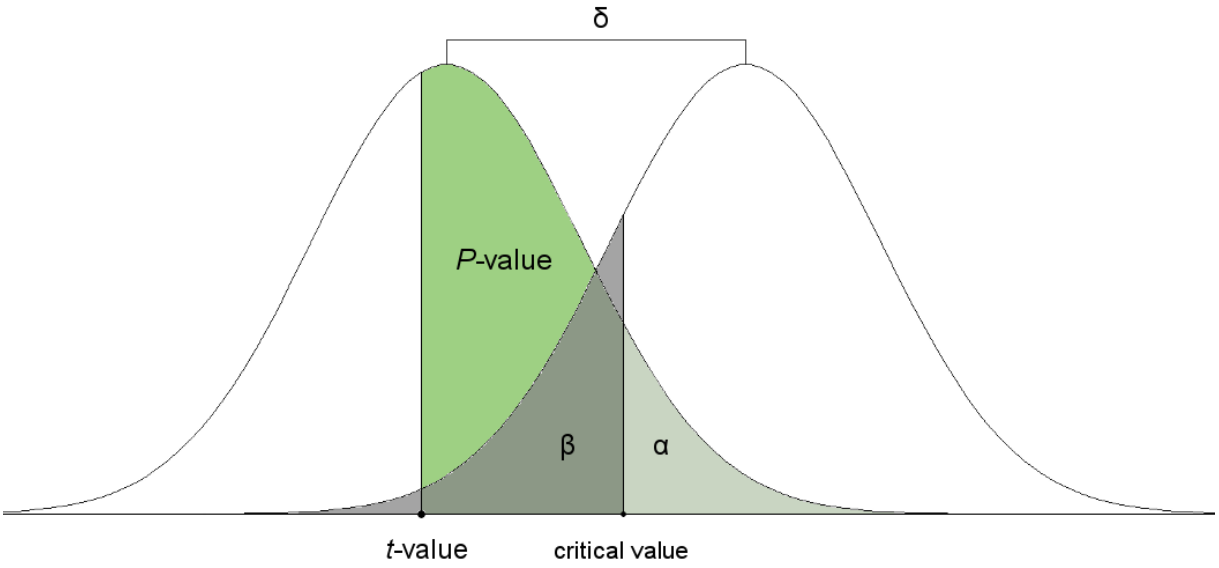


Figure 3

Density plot of observed (non)significant effect sizes for the eight selected journals, with 43% of effects in the category none-small, 27% small-medium, 12% medium-large, and 18% beyond large.

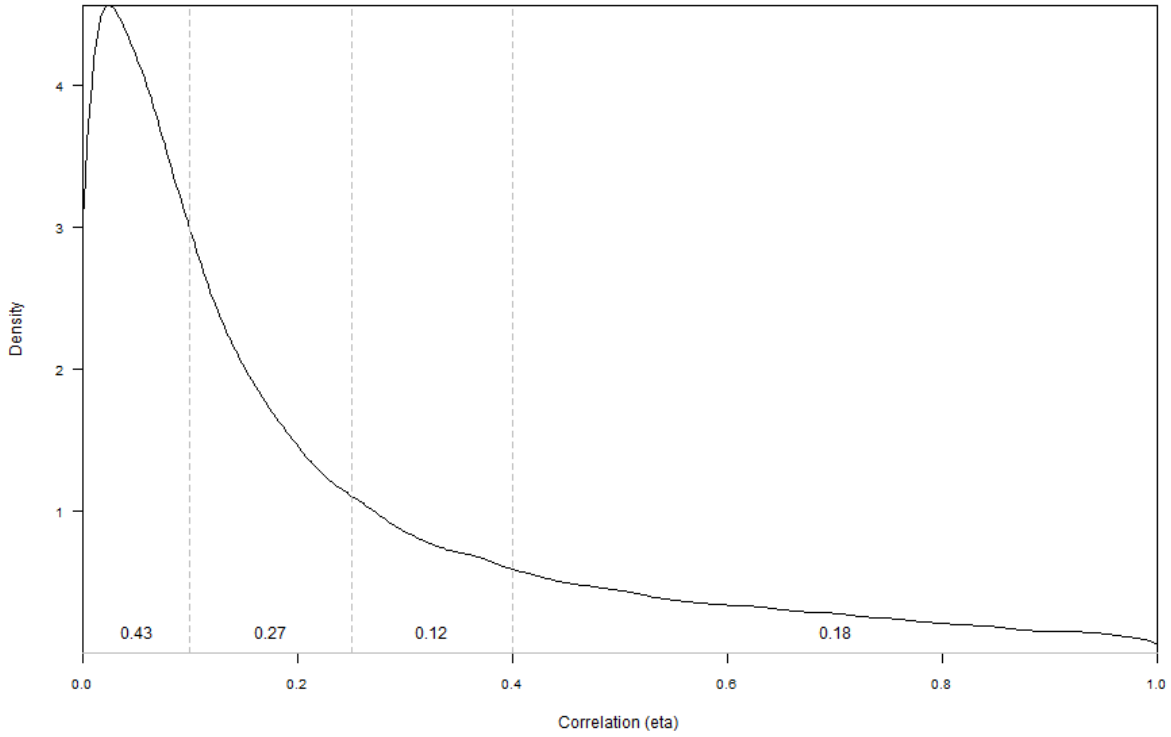


Figure 4

Observed proportion of nonsignificant test results per year.

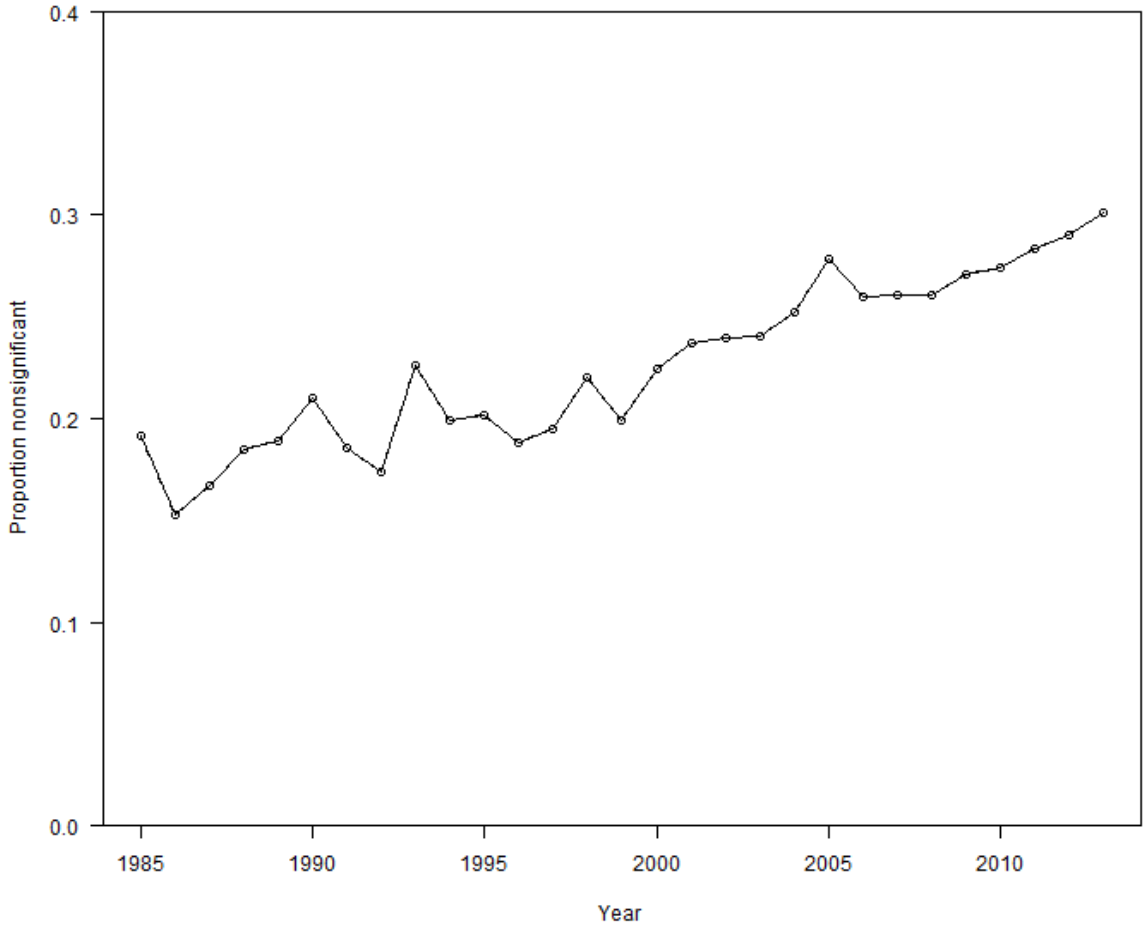


Figure 5

Observed effect distribution and expected no effect distribution for unadjusted- and adjusted effect sizes. Plot titles include Kolmogorov-Smirnov test results.

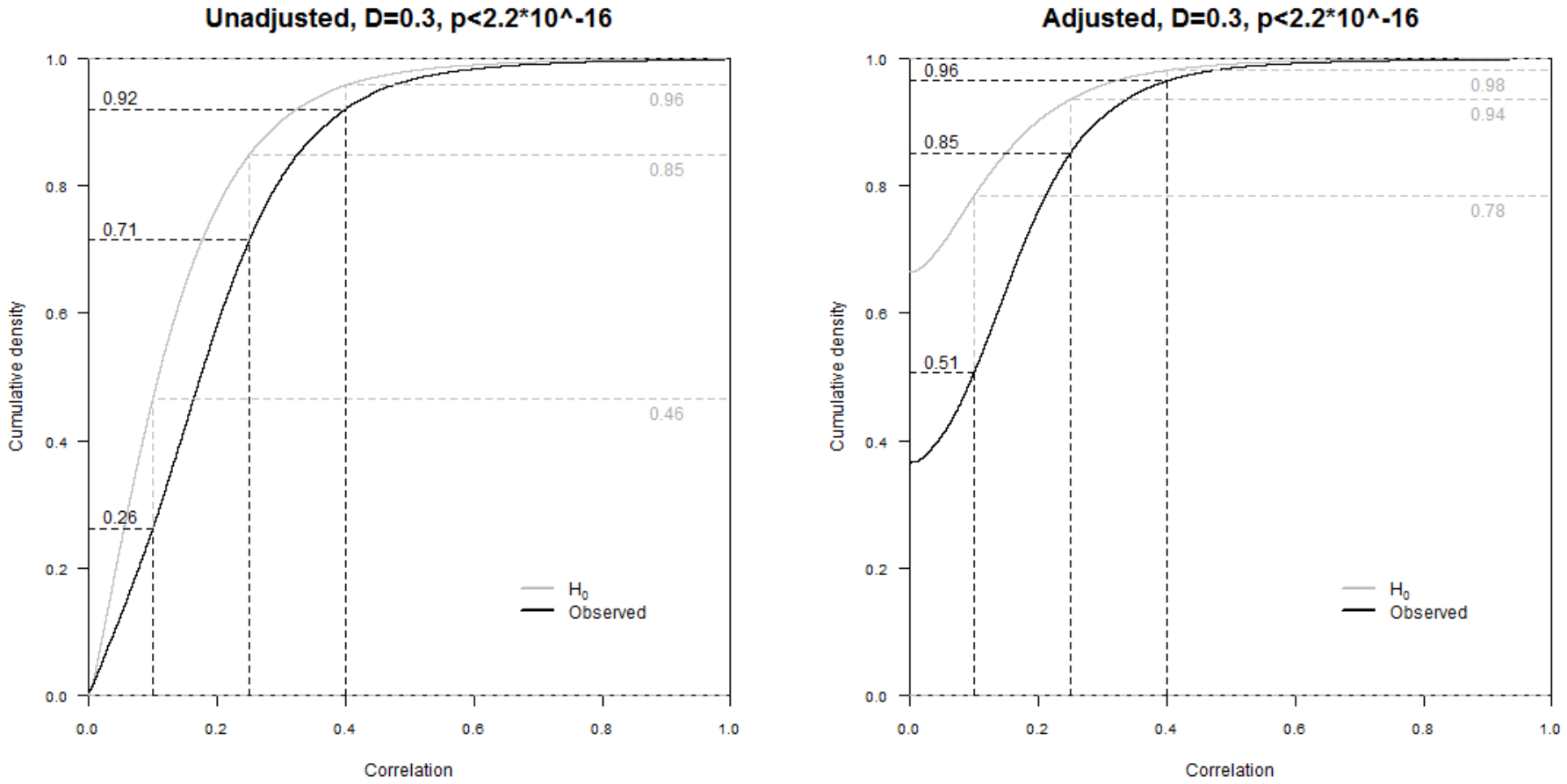
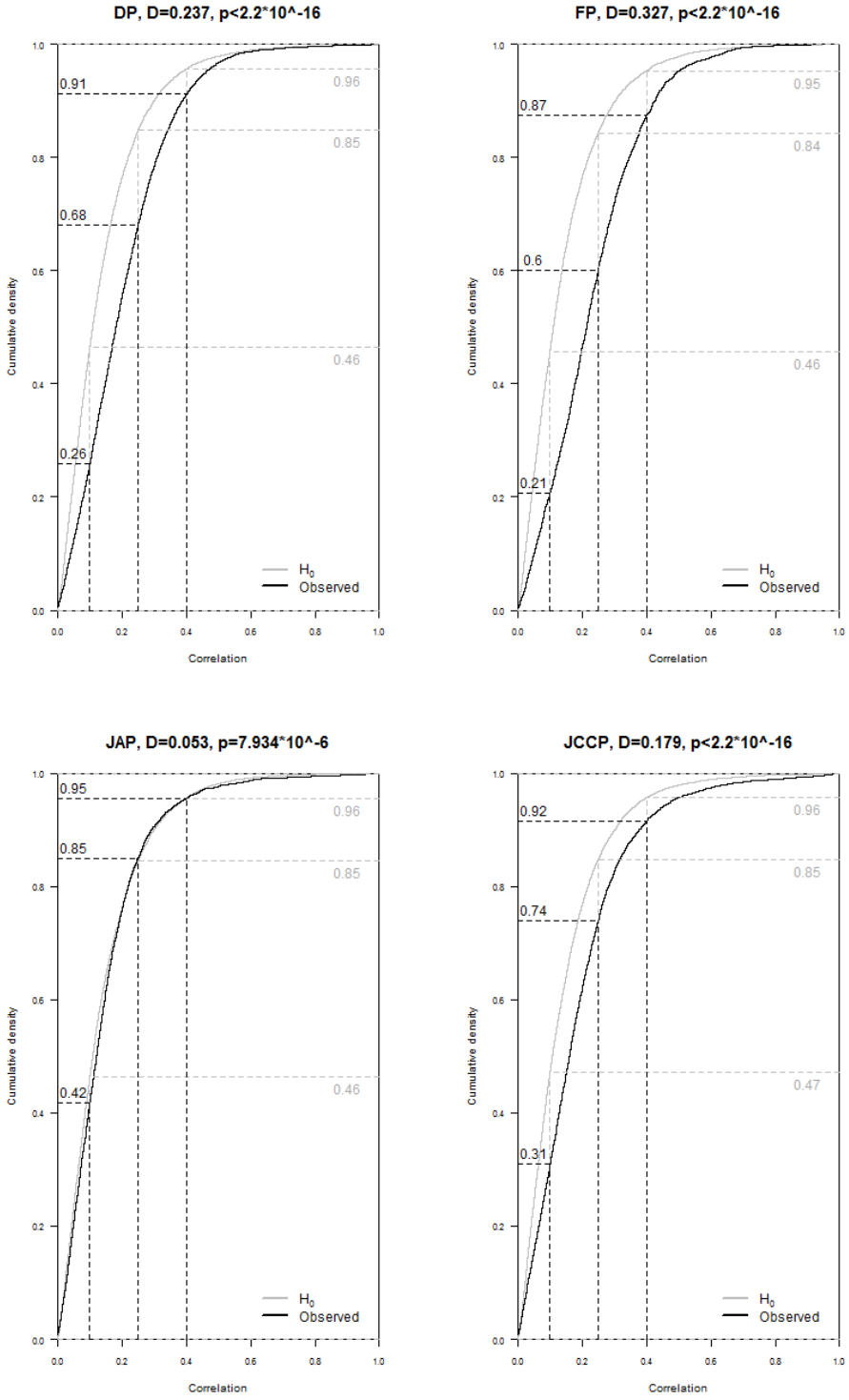
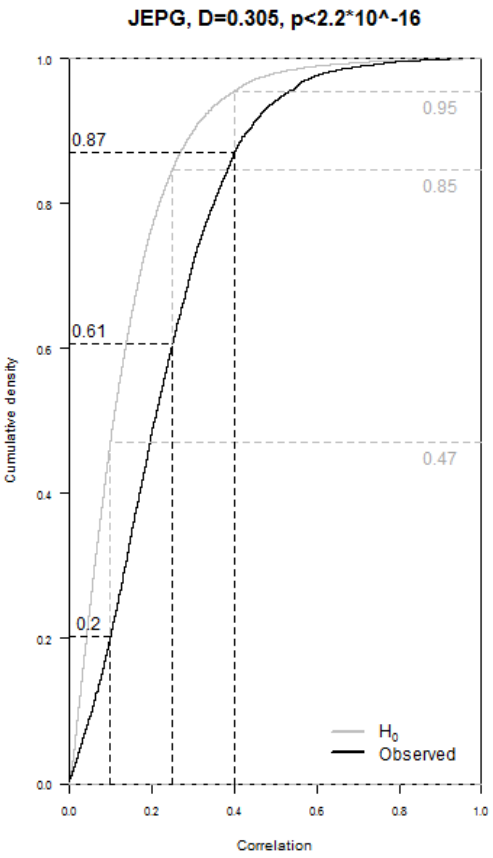
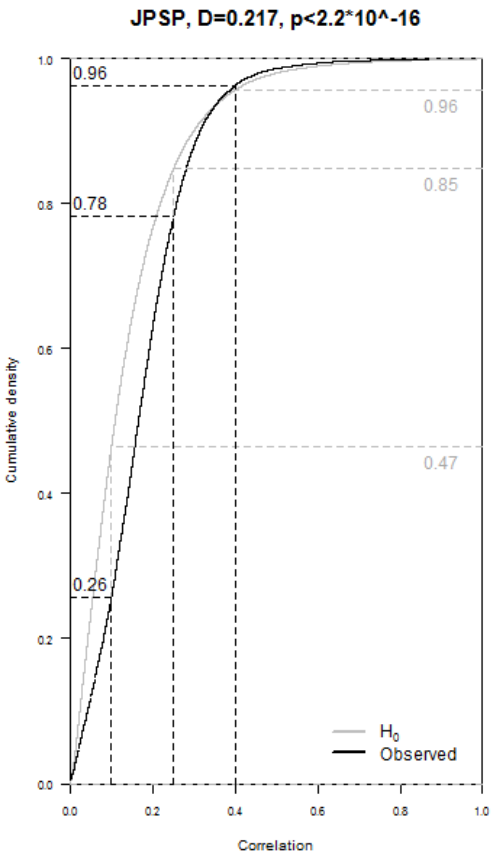


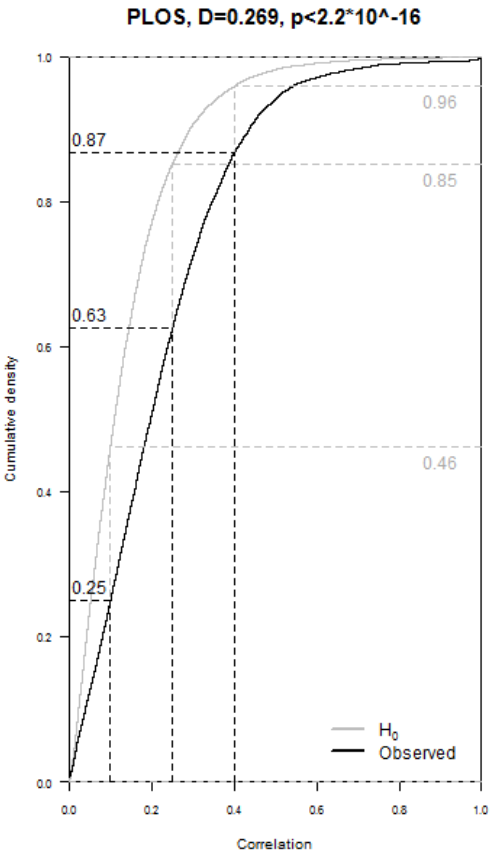
Figure 6

Observed- and expected nonsignificant effect distributions specified per journal. Plot titles include Kolmogorov-Smirnov test results. DP = Developmental Psychology; FP = Frontiers in Psychology; JAP = Journal of Applied Psychology; JCCP = Journal of Consulting and Clinical Psychology; JEPG = Journal of Experimental Psychology: General; JPSP = Journal of Personality and Social Psychology; PLOS = Public Library of Science; PS = Psychological Science.




JEPG, D=0.305, p<2.2*10^-16
Cumulative density
Correlation
0.87
0.61
0.2
0.95
0.85
0.47
H0
Observed


JPSP, D=0.217, p<2.2*10^-16
Cumulative density
Correlation
0.96
0.78
0.26
0.96
0.85
0.47
H0
Observed


PLOS, D=0.269, p<2.2*10^-16
Cumulative density
Correlation
0.87
0.63
0.25
0.96
0.85
0.46
H0
Observed

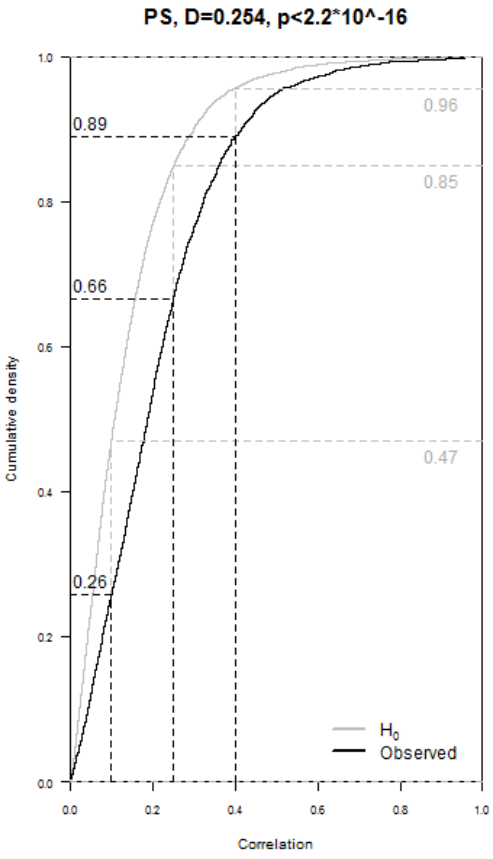

PS, D=0.254, p<2.2*10^-16
Cumulative density
Correlation
0.89
0.66
0.26
0.96
0.85
0.47
H0
Observed

Figure 7

False negative effect estimates, median ± 25 percentile points *N*, median- and mean *k* per year.

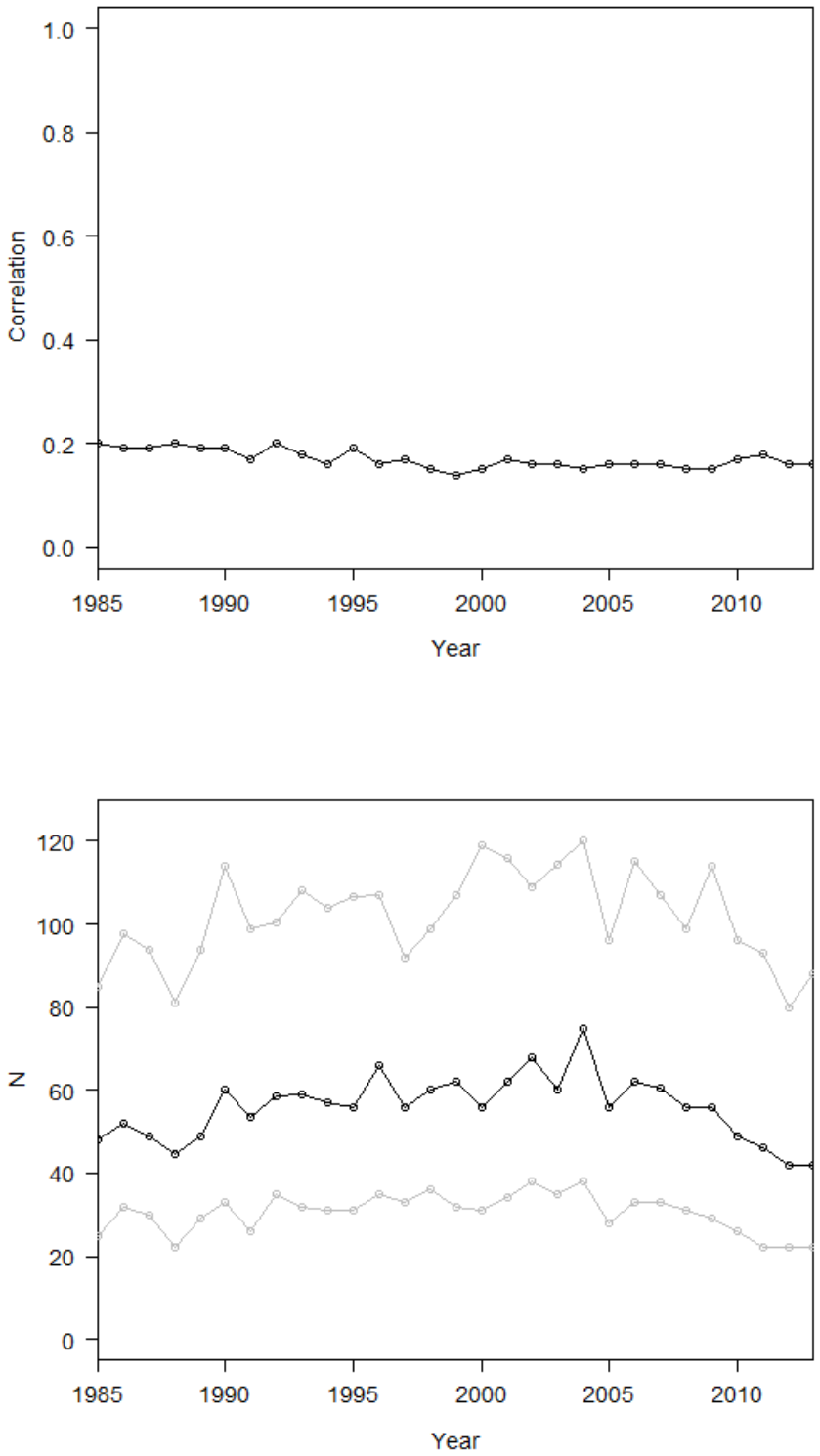


Figure 8

Estimated proportion false negatives per year.