Too Good to be False: Nonsignificant Results Revisited

Chris H.J. Hartgerink1

1 Tilburg University, the Netherlands

WORD COUNT: XXXX

Author note

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

Abstract

\*placeholder\*

*Keywords:* p-value, nhst, underpowered, effect size, fisher

**Too Good to be False: Nonsignificant Results Revisited**

Popper's (1959/2005) falsification 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, the bifurcated accept-reject decision becomes subject to decision errors, due to chance fluctuations. We review current findings on decision errors and investigate whether, and to what degree, there is evidence for these errors.

Null Hypothesis Significance Testing is the most prevalent statistical paradigm (NHST; American Psychological Association, 2010) and decision errors are recognized as an inherent theoretical property. In NHST, an ordinal point hypothesis is tested (i.e., *H0*), in which the point hypothesis regards the absence of an effect; a statement that is either true or false. If deemed false, an alternative, mutually exclusive hypothesis is accepted to provide a better depiction of reality (i.e., *H1*). Bifurcated decisions in NHST are based on the *P-*value: the probability of the sample data given that the point hypothesis is true (i.e., *P*[*D|H0*]). If the sample data is improbable beyond a certain decision criterion (i.e., α), this is regarded as proof that the point hypothesis is false. However, due to mere chance, we might decide in α of the samples that the data is proof of falsification, whereas in reality the point hypothesis is true (i.e., false positive). Conversely, there are cases where we lack the evidence to falsify the point hypothesis, while it is in fact false (i.e., false negative). Table 1 summarizes the four possible situations that are the consequence of bifurcated probabilistic decision making.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1 | | | |
|  |  | Population | |
|  |  | H0 | H1 |
| Sample | ‘H0’ | 1-α  [0.95] | β  [0.20] |
|  | *True negative* | *False negative* |
| ‘H1’ | α  [0.05] | 1-β  [0.80] |
|  | *False positive* | *True positive* |

After deciding whether a result is significant or not, error rates tend to be forgotten in assessing the results of a study. In other words, a form of error-blindness occurs with respect to chance findings. The NHST paradigm becomes problematic only in light of this forgetfulness. In any practical research situation the legitimacy of the findings remain unknown; hence it also remains unknown whether a decision is correct or false. Unconscious psychological fallacies, such as the belief in the law of small numbers (Tversky & Kahneman, 1971), bias against the null (Greenwald, 1975), or overconfidence in the true positive rate (Bakker, 2014; Greenwald, 1975) can cause unconscious underestimation of error-rates, increasing such forgetfulness. Additionally, errors might not be readily noticed across multiple studies, as humans have difficulty with judging properties of random variables deductively (Bar-Hillel & Wagenaar, 1991). This falsely increases confidence in the legitimacy of results and can even result in null findings blatantly being misinterpreted as the null hypothesis being true (Nickerson, 2000). This problem of misinterpretation and unconscious subjectivity in interpreting statistical results can lead to error-blindness, which gives a false sense of confidence in the accuracy of subjective conclusions.

This error-blindness has protruded the last three years, as discussion has caught on about false positive findings in the psychological sciences (e.g., Francis, 2012a, 2012b, 2014). (Un)successful replications have made the notion of false positives more salient, increasing awareness about the possibility of false positives. For example, the Many Labs Replication Project indicated that 2 out of 11 effects replicated were false positives (Klein et al., 2014) and that the seminal elderly priming study failed to replicate (Bargh, Chen, & Burrows, 1996; Harris, Coburn, Rohrer, & Pashler, 2013). Even though such replication attempts have been received critically (Bohannon, 2014), they do provide some indication of possible false positive findings. Additionally, widespread questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012) have been shown to have non-trivial effects on the chances of finding a false positive result (Simmons, Nelson, & Simonsohn, 2011). This concern has been addressed by the introduction of new methods to estimate whether findings are “too good to be true” (Francis, 2012; Ioannidis & Trikalinos, 2007; Schimmack, 2012), but also with rigorous methodological reforms (Asendorpf et al., 2013; Murayama, Pekrun, & Fiedler, 2013). These reforms and methods have indicated that the field of psychology is becoming less blind to false positive errors — at the paper level.

False positive rates also require inspection at a field level for an overview of the situation in the psychological sciences. The true positive rate (i.e., power) can be estimated based on the sample- and effect size, resulting in power estimates ranging between 35-50% for the field of psychological science (Bakker, Van Dijk, & Wicherts, 2012; Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Based on these power levels an estimate of the false positive rate can be made, given the number of significant findings (i.e., 95-97%; Sterling, Rosenbaum, & Weinkam, 1995; Sterling, 1959). Assuming that, before conducting a study, the alternative hypothesis is either three times less likely than the null (i.e., *R =* .25) or just as likely as the null (i.e., *R* = .5; Ioannidis, 2005), a conservative estimate of the false positive rate can be calculated, assuming alpha is .05.[[1]](#footnote-2) This estimate is calculated as the inverse of the positive predictive value (PPV), defined as 1 – [(1 – β)*R* / (1 *-* β*R* *+* α(1 – *R*))]. Given an *R* of .25 and 1-β of .35, or .5 and .5, respectively, the resulting estimate ranges between 9-30%, a number that qualifies the concerns.

Detecting false positives in the literature is more straightforward than detecting false negatives, maintaining error-blindness with respect to false negatives. This difference is primarily due to negative results being more readily interpreted as being conducted improperly or by an incompetent researcher (i.e., ad hominem fallacy), and negative results also have a lower probability of getting published. This is explained by a bias against null results (Greenwald, 1975). Such an initial bias against negative results, makes detecting a specific type of null results – the false negatives – especially difficult. Correction can only take place after detection has occurred, hence, correction of false negatives is also difficult (Fiedler, Kutzner, & Krueger, 2012). This is especially problematic, because false negative findings can cause valuable research lines to be stopped prematurely, foregoing a worthwhile investment, whereas false positives prolong research lines, continuing wasteful spending (Fiedler et al., 2012).

The false negative rate can also be estimated at the field level for the psychological sciences. This estimate is the inverse of the negative predictive value (NPV), defined as 1 – [β*R* / (β*R* *+* (1 – α)(1 – *R*))]. Given an *R* of .25 and β of .5, or .5 and .65, respectively, the estimated false negative rate ranges between 14.9-40.6% for the field of psychological science. This indicates that the estimated false negative rate, at a field level, is larger than the estimated false positive rate (i.e., 9-30%). False negative rates can be contained by increasing power, as false negative rate is defined as the inverse of power. Examples are power analysis and meta-analysis (Cumming, 2013). However, power analysis is troubled by misestimated effect sizes and meta-analysis is troubled by publication bias. These two methods have a non-recursive feedback cycle, where inflated effect sizes from meta-analyses, due to publication bias, yield erroneous power estimates, which consequently leads to new biased publication procedures. Unbiased effect estimation is under development (van Assen, van Aert, & Wicherts, 2013), but methods to inspect false negatives directly, on a paper level, are still absent.

The current paper investigates false negatives on both the field- and paper level, where the main question revolves around whether and to what degree there is evidence for false negative results in the published psychological sciences. To this end, nonsignificant results from eight flagship psychological journals were investigated. A six-pronged approach applied to this investigation, moving from ordinal, field level indications of false negatives to estimated, paper level indications of false negatives. After reviewing the theoretical properties applied throughout the paper, we first compare observed effect distributions to null distributions, where we expected the observed distributions to significantly deviate from the null distributions. This serves as an ordinal indicator of false negatives on a field level. Second, power simulations of the Fisher method were conducted. The Fisher method is a test that can indicate false negative findings on a paper level. Third, we applied the Fisher method to the observed data and report descriptive results. Fourth, the results from the Fisher method were used to estimate ad hoc effect sizes per journal. Fifth, the relation between the number of nonsignificant test results in a paper and the number of significant results from the Fisher method was inspected. Journals can differ in the amount of nonsignificant results reported, which needs to be taken into account when comparing results across journals. Sixth, use cases were selected on a standardized basis from the Journal of Personality and Social Psychology, to indicate the fruitfulness of the test with respect to practical research situations.

**Theoretical framework**

*P*-values are uniformly distributed when the null hypothesis is correct and the assumptions of the model are met. In case the null is false or these assumptions are violated, the *P*-value distribution is skewed. In other words, deviation from uniformity indicates either that the assumed null is not applicable in the sample data, or that model assumptions have been violated. This property of uniformity has been proposed to estimate effect sizes or QRPs in significant *P*-values (Simonsohn, Nelson, & Simmons, 2014; van Assen et al., 2013). The Fisher method (Fisher, 1932) tests deviation from uniformity and has been used as a meta-analytic technique (Hong & Breitling, 2008) to test for the presence of an effect across a set of studies. We use the property of uniformity to estimate a null-effect distribution and we propose that the Fisher method can be used to inspect for false negatives across nonsignificant *P*-values.

An individual *P-*value is a function of the sample size, the null hypothesis, and the observed effect size. This indicates that there is a direct relation between *P*-values and effect sizes, a property that can be used to compute the expected null distribution for effect sizes. For example, if the result of a *t*-test is *t*(85) = 2.15, p = .017, the accompanying standardized effect size *d* equals 0.46. Assuming the null hypothesis is correct, the *P-*values are uniformly distributed. Retaining only *t*(85) and assigning random drawn, uniform *P*-values, the accompanying test statistic and effect size can be computed. When conducted across a set of test results, comparing the observed- versus the expected null distribution is a way of testing for the presence of an effect.

Uniformity of *P*-values themselves can be tested with the Fisher method (Fisher, 1932), which only requires a set of *P*-values. This test is defined as



where *pi* is a vector of independent *P-*values, and *k* is the number of values in this vector. The resulting chi-square test statistic has twice the number of degrees of freedom as the number of *P*-values (i.e., 2*k*). Results of this test indicate the degree of deviation from uniformity; if significant there is evidence for deviation from the null. When only nonsignificant results are inspected, the set of nonsignificant *P*-values needs to be transformed back into the original [0; 1] state space, by computing adjusted *P*-values as



where *pi* is the vector of selected *P*-values, and α is the selected significance threshold. Note alpha is the significance threshold for the original test, not for the Fisher method. This transformation is required to retain the properties of NHST.

The uniformity of *P-*values under the null and the Fisher method will be used to investigate false negatives throughout the current paper. We apply these properties as described in the methods section.

**Methods**

**Procedure**

**Data summary.**The dataset was retrieved from the Open Science Framework[[2]](#footnote-3), and includes APA style test statistics extracted from 8 journals. These test statistics were extracted with statcheck (Epskamp & Nuijten, 2013) and originally included a total of XXXX test results (*t*, *r, F, Z,* χ2 and Wald values). As only *t*, *r,* and *F-*values allow for direct and comparable effect size computation, these test results were selected. For a more extensive description of the sampling method underlying the dataset, see the Open Science Framework page (Footnote 2).

**Effect computation.** The selected *t*, *F*, and *r*-values are readily computed into effect sizes, which form the observed effect distributions. The effect size metric used throughout the analyses is explained variance. For the selected *r*-values, this only requires taking the square (i.e., *r*2). Selected *F* and *t-*values are converted to effect sizes simultaneously, as squared *t-*values are *F-*values. The formula used to compute these effect sizes is



where, for squared *t*-values, *df1* equals 1, and *df2* equals the original degrees of freedom from the *t*-test. Adjusted effect sizes were computed as



and for *r*-values as



where the adjustment corrects for bias due to sample size.

Subsequently, these observed distributions were compared with the theoretical null distribution, with the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test 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 (overall, and per journal) with the null effect distribution was inspected. Differences in distributions between journals were not subjected to inferential significance tests, as the data are the population of *t*, *r,* and *F*-values reported in the journals. These procedures will indicate whether the observed nonsignificant effects differ from the null distribution, indicating false negatives, and whether there are differences in effects reported across different journals.

**Power simulations.** To simulate the power of the Fisher method, simulations were conducted across three factors (i.e., *N,* effect size, and the number of *P*-values). These factors were used to simulate *t*-test results, with 10,000 iterations for each condition. As mentioned, squared *t*-values are *F*-values. Consequently, results are readily generalizable to *F*-tests. We do not consider simulation results to be directly dependent on the number of groups, as power is a function of effect, *N*, and α — not group number.[[3]](#footnote-4) Levels for *N* were set at the 25th percentile, the 50th percentile (i.e., median), and the 75th percentile of the observed *df2* in the dataset (total: 3 levels). Effect sizes in the form of correlations were specified at 0.00 through 0.99, in 0.01 steps (i.e., 0.00, 0.01, 0.02, 0.03, …, 0.99; total: 100 levels). The number of *P*-values was specified at 1 (total: 18 levels). Per condition, 10,000 iterations were run, where each iteration yielded a *P*-value for the Fisher test. The proportion of significant iterations (α = .10) indicates sensitivity of the test for effect size zero, and power for all other effect sizes. If the resulting power in one condition was 99.5% or higher, power for subsequent effect sizes was automatically overridden to be 1 (given *N* and number of results), because larger effect sizes result in higher power (Aberson, 2010). This was done to counteract increases in runtime and decreases in precision for large non-centrality parameters (Lenth, 1989).

For each condition, 9 steps were necessary to simulate the result of the Fisher method. First, a critical value under the null distribution was needed (α = .05), which was readily computed by using the degrees of freedom (i.e., *N*-1). Second, the degrees of freedom and effect size were used to compute the non-centrality parameter. The non-centrality parameter (i.e., δ) is computed as (Smithson, 2001; Steiger & Fouladi, 1997)



where



This non-centrality parameter adjusts the null distribution for the population effect (e.g., shifts it to the right, see Figure 1). Third, for the population distribution, the area under the curve for a nonsignificant result was determined (i.e., β). Fourth, a value was uniformly drawn between 0 and the β value that resulted from step three. Fifth, the accompanying *t*-test-value was computed, which was, sixth, used to compute the *P-*value under the null distribution. Sixth, this nonsignificant *P*-value was transformed with Equation 3. Seventh, for each set of *P*-values per condition per iteration, a Fisher test was conducted and the resulting *P*-value saved. Finally, sensitivity/power was calculated as the proportion of significant results (α = 0.10).

**Effect estimation.** The simulation procedure can be used on observed results to estimate an effect size. Given a hypothetical population effect size, power of the Fisher method can be computed across all the individual papers in the dataset. Summing the power values for each paper, given an effect size, results in the expected value of significant Fisher tests for that effect size. Subsequently, the observed number of significant Fisher tests can be compared with the expected number of significant Fisher tests for certain effect sizes. The effect size that minimizes the discrepancy between the observed and expected proportion of significant Fisher method tests subsequently provides an ad hoc effect size estimate.

Additionally, the relation between the number of significant Fisher method results can be related to the number of test results in a paper.

**Use case.**

**Results**

**Effect distribution**

***P*-value distribution**

**Power simulations.**

**Effect estimation.**

**Use case.**

**Discussion**

References

Footnotes

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Population | |
|  |  | H0 | H1 |
| Sample | ‘H0’ | 1-α  [0.95] | β  [0.20] |
|  | *True negative* | *Type II error* |
| ‘H1’ | α  [0.05] | 1-β  [0.80] |
|  | *Type I error* | *True positive* |

*Note.* Columns indicate the true situation in the population, rows indicate the statistical conclusion based on sample data. The true positive rate is also called power, and the true negative rate is also called XXXX. Values in square brackets are conventionally acceptable values.

Table 2

Table 3

*Note.* Alpha was assumed to be .05 to determine significant or not.

*Figure 1*

Visual depiction of steps 1 through 6 for simulation procedure of Fisher tests.

*Figure 2*

Observed effects versus simulated null effects.

*Figure 3*

Plots of the power (y-axis) across simulation results. Dashed lines indicate medium and large effect sizes, respectively.

1. This implicitly assumes no QRPs, hence, is a conservative estimate, because QRPs increase the number of false positives. [↑](#footnote-ref-2)
2. <https://osf.io/dzrtf/> [↑](#footnote-ref-3)
3. More groups does decrease power indirectly, because of decreased sample size per group, keeping all else constant. [↑](#footnote-ref-4)