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. Within the theoretical framework of scientific hypothesis testing, accepting or rejecting a hypothesis is unequivocal, because the hypothesis is either true or false. Such a bifurcated decision is warranted, but only on a theoretical level. 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 is unwarranted. Any decision regarding the hypothesis is possibly erroneous merely due to chance fluctuations. These chance fluctuations are regularly and unconsciously neglected when deciding whether a hypothesis is falsified or not — causing errors to propagate in the academic literature.

This propagation of decision errors is partially due to the most commonly used statistical paradigm: Null Hypothesis Significance Testing (NHST; American Psychological Association, 2010). 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.[[1]](#footnote-2) 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* |

The NHST paradigm becomes problematic only in light of the people using it; misinterpretation and unconscious psychological fallacies undermine the theoretical properties. It is only this combination that warrants concern. In any practical research situation the legitimacy of the findings remain unknown; hence it also remains unknown whether a conclusion 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. 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 (Francis, 2012a, 2012b, 2014). Unsuccessful replications of findings has taken the idea of false positives out of the abstract and into the concrete, increasing concern. For example, the seminal elderly priming study failed to replicate (Bargh, Chen, & Burrows, 1996; Harris, Coburn, Rohrer, & Pashler, 2013). 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” (i.e., false positives; 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 need to be inspected at a field level to get a better overview of the situation in the psychological sciences in general. Positive findings in the field of psychological science, including both true- and false positives, have been estimated at 95-97% (Sterling, Rosenbaum, & Weinkam, 1995; Sterling, 1959). 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). Under the stringent assumptions that the null hypothesis and alternative hypothesis are equally likely to be true (i.e., *R* = .5; Ioannidis, 2005), that there is no publication bias, and QRPs are absent, a conservative estimate of the false positive rate for significant findings can be calculated to range between 20-27.5%,[[2]](#footnote-3) 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 conducted by an incompetent researcher (i.e., ad hominem fallacy), but also have a lower probability of getting published. These are both explained by a bias against null results (Greenwald, 1975). Such an initial bias against null 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).

Estimating the false negative rate, based on the inverse of the procedure to estimate false positives (i.e., null and alternative are equally probable, no publication bias, no QRPs[[3]](#footnote-4)) yields a false negative rate ranging between 72.5-80% for the field of psychological science. This indicates that the problem of false negatives, at a field level, is possibly larger in size than that of false positives, and would, at least theoretically, be more difficult to tackle when compared to the problem of false positives. Methods to decrease the false negative rate on a paper level relate directly to methods to increase the power of studies, as false negative rate is defined as the inverse of power. Examples are power analysis and meta-analysis (Cumming, 2013), but problems with practical implementations of these methods makes easier and readily implementable methods warranted. Power analysis is troubled by misestimating the effect size and meta-analysis is troubled by publication bias. These two methods have a non-recursive feedback cycle, where inflated effect sizes from meta-analyses yield erroneous power estimates, which consequently leads to possibly biased publication procedures. Unbiased effect estimation is under development (van Assen, van Aert, & Wicherts, 2013), but methods to inspect false negatives directly 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 five-pronged approach applied to this investigation, moving from ordinal, field level indications of false negatives to estimated, paper level indications of false negatives.

First, observed effect distributions were computed based on the nonsignificant test results reported in the journals. These distributions were compared to the null distributions. We expected the observed distributions to significantly deviate from the null distributions, which serves as an ordinal indicator of false negatives.

Second, the Fisher method (explained below) is proposed as a succinct and easy-to-use test for detecting possible false negatives. This test was simulated across a series of conditions to inspect the power of the method. We expected power to increase as a function of the factorial levels sample size, effect size, and the number of test results (Aberson, 2010) and were primarily interested in getting numeric estimates of power levels across a range of conditions.

Third, the simulation procedure was applied to the observed test results, to estimate the power of finding a false negative on a paper level, given imposed, hypothetical effect sizes. Combined with the results of the Fisher method per paper, this allows for ad hoc estimation of effect sizes across the field of psychological sciences and for journals specifically. We expected the ad hoc effect estimates to show the presence of an effect (i.e., > 0) and provide a crude indication of the degree of the effect ‘hidden’ in the nonsignificant results.

Fourth, a curve fitting procedure was applied to estimate statistical power of the Fisher method. This curve is a function of the number of test results presented in a paper and an effect size, and can provide ad hoc power estimate given the effect. Such an ad hoc estimate can be used to retrospectively compute the number of true positive results, and the number of false negative results, of the Fisher method.

Fifth, several 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**

Statistical hypothesis testing in the form of NHST is a theoretical framework with one crucial property: *P*-values are uniformly distributed under the null distribution. This uniformity of *P*-values is a theoretical principle and is easily shown in simulations (Murdoch, Tsai, & Adcock, 2008; Sackrowitz & Samuel-Cahn, 1999). Conversely, when calibrating a new method, uniformity under the null is an important element to ensure specificity of a test. In case the null is false, the *P*-value distribution is skewed. Deviation from uniformity has been proposed to estimate effect sizes (left skewed) or QRPs (right skewed) in significant *P*-values (Simonsohn, Nelson, & Simmons, 2014; van Assen et al., 2013). We propose it can also be used to inspect for false negatives across nonsignificant *P*-values.

This property of uniformity can be used (1) to estimate null effect distributions and (2) to test whether observed *P-*values deviate from uniformity. Null effect distributions can be computed by drawing uniformly distributed *P-*values and computing accompanying test values, which can then be used to compute effect sizes. Uniformity of *P*-values themselves is 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. Related tests, which have been previously applied to test publication bias, are typically used with an alpha level of 10% (e.g., Francis, 2012b; Ioannidis & Trikalinos, 2007), and equivalent alpha will be used for the Fisher method throughout this paper.

In the current paper, we only inspect whether nonsignificant *P*-values deviate from uniformity. As selecting nonsignificant *P*-values imposes range restriction, the selected *P*-values are transformed back into the original state space of [0; 1], by computing adjusted *P*-values as



where *pi* is the vector of selected *P*-values, and α is the selected significance threshold. Note this alpha is the significance threshold for the original test, not for the Fisher method, and was assumed to be 5% across all test results. This transformation is required to retain the properties of NHST that underlie the current paper.

**Methods**

**Procedure**

***Data summary.*** The dataset was retrieved from the Open Science Framework[[4]](#footnote-5), 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 (XXXX results; XX% of original). Table 2 summarizes the selected data used for the analyses in this paper. For a more extensive description of the sampling method underlying the dataset, see the Open Science Framework page (Footnote 2).

**Effect distribution**

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



where, for squared *t*-values, *df1* equals 1, and *df2* equals the original degrees of freedom from the *t*-test. Adjusted effect sizes are also computed (see Appendix for adjusted formula).

The theoretical null distribution across all test results was simulated. This was done by (1) randomly sampling 1,000,000 nonsignificant test results from the dataset with replacement, (2) sampling nonsignificant *P*-values uniformly between 0 and 0.95 (i.e., α = .05), representing the null, and (3) computing the effect size that accompanies the degrees of freedom for the simulated test results. Effect size computation is the same as computing the observed effect sizes, except no adjusted effect sizes are computed.

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, because of which 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.[[5]](#footnote-6)

Factorial specification resulted in 1600 conditions. Levels for *N* were set at 25, 50, 100, and 150 (total: 4 levels). Effect sizes in the form of eta-squared were specified at 0.00 and 0.01 through 0.95, in 0.01 steps (i.e., 0.01, 0.02, 0.03, …, 0.95; total: 100 levels). The number of *P*-values was specified at 2, 4, 7 and 10 (total: 4 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. Power estimates Summing the power values for each test statistic in the dataset, 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. Depending on the precision of the simulations, with regards to effect sizes, an effect estimate can be made in a certain

**Use case.**

**Results**

**Effect distribution**

***P*-value distribution**

**Power simulations.** Simulations indicated that the Fisher method is highly powerful when ordinally testing for the presence of an effect. Figure 3 visually summarizes the results, and clearly indicates that power increases as a function of the factors. In other words, the power of the Fisher method increases as *N* per individual result increases, as the number of results increases, and as the effect size increases. This is in line with power theory on an individual test level (Aberson, 2010).

**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 point hypothesis is commonly 0 (i.e., nil hypothesis), but can be any point value. [↑](#footnote-ref-2)
2. Calculated as: *R* × α + (1 – *R*) × (1 – β), i.e., .5 × .05 + (1 – .5) × [.35, .50] [↑](#footnote-ref-3)
3. It is noteworthy that, as QRPs become more prevalent, power will increase and the false negative rate will decrease, whereas the false positive rate will increase. This is not a 1:1 relation, as power increases will slow down the larger power gets. [↑](#footnote-ref-4)
4. <https://osf.io/dzrtf/> [↑](#footnote-ref-5)
5. More groups does decrease power indirectly, because of decreased sample size per group, keeping all else constant. [↑](#footnote-ref-6)