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) falsificationism has become deeply rooted in current day philosophy of science, which, in tandem, has caused hypothesis testing to become ubiquitous in the social sciences. Falsification serves as one of the main demarcating criteria in the social sciences, upon which the theoretical logic of scientific hypothesis testing is founded. Statistical hypothesis testing is an operationalization of scientific hypothesis testing, but these diverge in their logical rigor. Whereas outcomes of theoretical scientific hypothesis tests are deterministic, its operationalization makes outcomes probabilistic.

In quantitative research, statisticalhypothesis tests are often used. The most common statisticalhypothesis tests are Null Hypothesis Significance Tests (NHST), where a point hypothesis[[1]](#footnote-2) is tested against an alternative hypothesis. The probabilistic nature of specifically these types of statistical hypothesis tests yields individual results that are (1) inconclusive, (2) incomplete, and (3) error-prone. Inconclusive results can be seen in structural equation modeling, where different models yield identical results (Kline, 2011). Results are incomplete, as a *P*-value does not state the probability of a hypothesis (Cohen, 1994), and results are error-prone due to sampling fluctuations. Table 1 gives an overview of possible results based on NHST. These connotations do not invalidate the theoretical model of statistical hypothesis testing, but require attention as results are often misunderstood or misinterpreted (for a comprehensive review see Harlow, Mulaik, & Steiger, 1997). The errors that can result from statistical hypothesis testing are the focus of this paper, and can be categorized into false positives and false negatives. More specifically, the paper pertains to the prevalence rates of such errors.

False positives have been widely discussed in recent years, as events and papers have clearly indicated the plausibility of widespread false positive findings in the literature. Extreme and (hopefully) incidental fraud cases (e.g., Hauser, Stapel, Smeesters, Sanna, Ruggiero, and Förster) presented clearly false positive results; papers have also indicated that widespread questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012) can have dramatic effects on false positive rates (Simmons, Nelson, & Simonsohn, 2011). Questionable editorial practices (QEPs; LeBel et al., 2013), publication bias (Francis, 2012a, 2013), systemic low power (Bakker, Van Dijk, & Wicherts, 2012), and excessive degrees of significance (Francis, 2012b, 2014; Schimmack, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995) are additional factors that could have led to high false positive rates in the academic literature. Failures to replicate effects, such as the elderly priming effect (original: Bargh, Chen, & Burrows, 1996; replication: Harris, Coburn, Rohrer, & Pashler, 2013), physical and moral cleansing effect (original: Zhong & Liljenquist, 2006; replications: Earp, Everett, Madva, & Hamlin, 2014; Fayard, Bernstein, & Roberts, 2009), and the ESP effect (original: Bem, 2011; replication: Ritchie, Wiseman, & French, 2012) cast serious doubt on the truth of these effects (e.g., Yong, 2012). Additionally, the need for innovative results in journals (Trafimow, 2013) could inflate the false positive rate in the literature further. It should be noted that the field of psychology took up these concerns, is open to changing its practices (Fuchs, Jenny, & Fiedler, 2012), is investigating remedies (e.g., Murayama, Pekrun, & Fiedler, 2013), and is implementing new standards for research practices (e.g., Brandt et al., 2013; Eich, 2013; Journal of Experimental Social Psychology, 2014).

False negatives have been discussed less explicitly, as the problems for false positives were more salient and more pressing. Nonetheless, power (i.e., true positive rate) has implicitly been part of the discussion, because replication requirements were being discussed for detecting possible false positives (e.g., Asendorpf et al., 2013; Brandt et al., 2013). Power is of importance, because the power for psychological studies has been estimated to be as low as 35% (Bakker et al., 2012), which implies a 65% false negative rate. This is substantial, given that this exceeds all but one of the false positive rates due to QRPs (81.5% in Table 1; Simmons et al., 2011). Systemic low power in psychology has been indicated before (e.g., Cohen, 1962), but this has had almost no effect on increasing power in the field (Sedlmeier & Gigerenzer, 1989). Discussing false negatives is of importance, because such false negatives are less easily discovered than false positives and causes research lines to be stopped prematurely, whereas false positives prolong research lines (Fiedler, Kutzner, & Krueger, 2012). Research has focused on the prevalence rate of false positives; less is known about the prevalence rate of false negatives.

This paper inspects the prevalence of false negatives, by focusing on null findings from statistical hypothesis tests. Null findings are often interpreted as the null hypothesis being true (Nickerson, 2000), which could be caused by unconscious psychological fallacies such as the law of small numbers (Tversky & Kahneman, 1971) or overconfidence about the power of a study (Bakker, 2014; Greenwald, 1975). Such fallacies increase the likelihood that null results are interpreted as true negatives, which is why we hypothesize to find widespread indication for false negative results in the literature. If null results are true negatives, *P*-values are uniformly distributed (Murdoch, Tsai, & Adcock, 2008; Sackrowitz & Samuel-Cahn, 1999). Our research investigates whether reported statistics from 8 flagship journals corroborates null distributions, or indicates possible false negative results.

**Methods**

**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 (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.

***P*-value distribution**

Uniformity of *P*-values is tested with the Fisher method (Fisher, 1932/1948), 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. 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 inspect whether nonsignificant *P*-values deviate from uniformity. As selecting nonsignificant *P*-values restricts the range, the selected *P*-values are transformed back into the original state space of [0; 1], by computing adjusted *P*-values with



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.

**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.[[3]](#footnote-4)

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

Aberson, C. L. (2010). *Applied Power Analysis for the Behavioral Sciences*. New York, NY: Routledge.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119. doi:10.1002/per.1919

Bakker, M. (2014). Flawed intuitions about power in psychological research. In *Good Science, Bad Science: Questioning Research Practices in Psychological Science [Dissertation]*.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi:10.1177/1745691612459060

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*(2), 230–244. doi:10.1037/0022-3514.71.2.230

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–25. doi:10.1037/a0021524

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153. Retrieved from http://psycnet.apa.org/journals/abn/65/3/145.pdf&productCode=pa

Cohen, J. (1994). The Earth is Round (p < .05). *American Psychologist*, *49*(12), 997–1003.

Earp, B. D., Everett, J. a. C., Madva, E. N., & Hamlin, J. K. (2014). Out, Damned Spot: Can the “Macbeth Effect” Be Replicated? *Basic and Applied Social Psychology*, *36*(1), 91–98. doi:10.1080/01973533.2013.856792

Eich, E. (2013). Business Not as Usual. *Psychological Science*, *25*(1), 3–6. doi:10.1177/0956797613512465

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

Fayard, J. V, Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness ? Dispelling old wives’ tales : Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis*, *6*(2), 21–30.

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*(6), 661–669. doi:10.1177/1745691612462587

Fisher, R. A. (1932/1948). *Statistical Methods for Research Workers* (10th ed.). Edinburgh, United Kingdom: Oliver and Boyd.

Francis, G. (2012a). Publication bias and the failure of replication in experimental psychology, 975–991. doi:10.3758/s13423-012-0322-y

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

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2013.02.003

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0601-x

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists Are Open to Change, yet Wary of Rules. *Perspectives on Psychological Science*, *7*(6), 639–642. doi:10.1177/1745691612459521

Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin*, *82*(1), 1–20.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum.

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

Ioannidis, J. P., & Trikalinos, T. a. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. doi:10.1177/1740774507079441

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 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/

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (Third.). New York, NY: The Guilford Press.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science*, *8*(4), 424–432. doi:10.1177/1745691613491437

Lenth, R. V. (1989). Algorithm AS 243: Cumulative Distribution Function of the Non-Central t Distribution. *Journal of the Royal Statistical Society*, *38*(1), 185–189.

Massey Jr., F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research Practices That Can Prevent an Inflation of False-Positive Rates. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*. doi:10.1177/1088868313496330

Murdoch, D., Tsai, Y., & Adcock, J. (2008). P-values are random variables. *The American Statistician*, *62*, 242–245.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10937333

Popper, K. (1959/2005). *The logic of scientific discovery*. London, United Kingdom: Routledge. Retrieved from http://books.google.com/books?hl=en&lr=&id=LWSBAgAAQBAJ&oi=fnd&pg=PP1&dq=The+Logic+of+Scientific+Discovery&ots=pxLk-Z5LcQ&sig=vzfk74fyPhOrjhWmkqk3U1S\_1xg

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: three unsuccessful attempts to replicate Bem’s “retroactive facilitation of recall” effect. *PloS One*, *7*(3), e33423. doi:10.1371/journal.pone.0033423

Sackrowitz, H., & Samuel-Cahn, E. (1999). P values as random variables—expected P values. *The American Statistician*, *53*(4), 326–331. Retrieved from http://amstat.tandfonline.com/doi/full/10.1080/00031305.1999.10474484

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 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*(2), 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*(11), 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*(4), 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), pp. 30–34. Retrieved from http://www.jstor.org/stable/2282137

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), pp. 108–112. Retrieved from http://www.jstor.org/stable/2684823

Trafimow, D. (2013). The importance of surprising findings in psychology. *Frontiers in Psychology*, *4*, 173. doi:10.3389/fpsyg.2013.00173

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. doi:10.1037/h0031322

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature*. doi:10.1038/nature.2012.11535

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: threatened morality and physical cleansing. *Science*, *313*(5792), 1451–2. doi:10.1126/science.1130726

Footnotes

Table 1

|  |  |  |
| --- | --- | --- |
|  | H0 | H1 |
| ‘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.

C:\Users\Chris\Dropbox\CJM\Masterproject\Writing\Figures\fig1.tiff

*Figure 2*

Observed effects versus simulated null effects.

C:\Users\Chris\Dropbox\CJM\Masterproject\Analyzing\b.Figures\null.tiff

*Figure 3*

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

C:\Users\Chris\Dropbox\CJM\Masterproject\Analyzing\b.Figures\sim.tiff

1. This point hypothesis is commonly 0 (i.e., nil hypothesis), but can be any point value. [↑](#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)