Empirical Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(3) 357–366

DOI: 10.1177/2515245918773742
www.psychologicalscience.org/AMPPS

Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation

Balazs Aczel[1], Bence Palfi[2,3], Aba Szollosi[4], Marton Kovacs[1], Barnabas Szaszi[1,5], Peter Szecsi[1], Mark Zrubka[1], Quentin F. Gronau[6], Don van den Bergh[6], and Eric-Jan Wagenmakers[6]

[1]Institute of Psychology, ELTE Eötvös Loránd University; [2]School of Psychology, University of Sussex; [3]Sackler Centre for Consciousness Science, University of Sussex; [4]School of Psychology, University of New South Wales; [5]Doctoral School of Psychology, ELTE Eötvös Loránd University; and [6]Department of Psychology, University of Amsterdam

Abstract

In the traditional statistical framework, nonsignificant results leave researchers in a state of suspended disbelief. In this study, we examined, empirically, the treatment and evidential impact of nonsignificant results. Our specific goals were twofold: to explore how psychologists interpret and communicate nonsignificant results and to assess how much these results constitute evidence in favor of the null hypothesis. First, we examined all nonsignificant findings mentioned in the abstracts of the 2015 volumes of *Psychonomic Bulletin & Review, Journal of Experimental Psychology: General*, and *Psychological Science* ($N = 137$). In 72% of these cases, nonsignificant results were misinterpreted, in that the authors inferred that the effect was absent. Second, a Bayes factor reanalysis revealed that fewer than 5% of the nonsignificant findings provided strong evidence (i.e., $BF_{01} > 10$) in favor of the null hypothesis over the alternative hypothesis. We recommend that researchers expand their statistical tool kit in order to correctly interpret nonsignificant results and to be able to evaluate the evidence for and against the null hypothesis.

Keywords

nonsignificant results, NHST, Bayes factor analysis, open data, open materials, preregistered

Received 12/20/17; Revision accepted 4/3/18

> Never use the unfortunate expression "accept the null hypothesis."
>
> —Wilkinson and the Task Force on Statistical Inference (1999, p. 599)

The interpretation of statistically nonsignificant findings is a vexing point of traditional psychological research.[1] Within the framework of null-hypothesis significance testing (NHST; Fisher, 1925; Neyman & Pearson, 1933), decisions about the null hypothesis are based on the p value. Under NHST logic, one is entitled to reject the null hypothesis whenever the p value is smaller than or equal to a predefined α threshold (typically set at .05; but see Benjamin et al., 2018). In contrast, the p value does not entitle one to claim support in favor of the null hypothesis. According to the common interpretation, any p value higher than α indicates that one has to withhold judgment about the null hypothesis (Cohen, 1994). This asymmetric characteristic of the NHST framework frustrates the interpretation and communication of nonsignificant results (Edwards, Lindman, & Savage, 1963; Nickerson, 2000). It is known that results with a p value greater than .05 are subject to misinterpretation among researchers (Goodman, 2008),

Corresponding Author:
E.-J. Wagenmakers, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018XE Amsterdam, The Netherlands
E-mail: ej.wagenmakers@gmail.com

but the extent to which this biases the communication of psychological findings has been unexplored. In this article, we examine the degree of miscommunication of nonsignificant findings in current psychological publications; in addition, we use Bayes factors to assess how much these findings support the null hypothesis relative to a composite alternative hypothesis (e.g., Etz & Vandekerckhove, 2017).

Nonsignificant findings in psychological research are both disliked and misinterpreted, and this brings dire consequences. First, the common aversion to nonsignificant findings (e.g., Ferguson & Heene, 2012; Greenwald, 1975) not only causes publication bias (e.g., Franco, Malhotra, & Simonovits, 2014) but also harms the validity of the reported outcomes. For example, most questionable research practices are aimed at transforming otherwise nonsignificant *p* values into significant *p* values (e.g., Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016; Lilienfeld & Waldman, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Pritschet, Powell, & Horne, 2016). Second, nonsignificant findings are commonly misinterpreted, usually because researchers regard nonsignificant *p* values as support for the null hypothesis (i.e., misconception #2 in Goodman, 2008). However, *p* values larger than the threshold indicate only that the test was incapable of rejecting the null hypothesis; this could have occurred because the effect does not exist, but it also could have occurred because the power of the test was insufficient to detect a true effect (Dienes, 2014, 2016). Indeed, an examination of the psychology literature suggests that a high proportion of statistically nonsignificant results are false negatives (Hartgerink, Wicherts, & van Assen, 2017).

Finally, when confronted with nonsignificant findings, researchers may seek refuge in a description of the sample rather than inference concerning the population; such a tendency is revealed by expressions such as "no difference between the groups was observed." Such statements about the sample are problematic, as the observed difference is never exactly zero in the case of continuous data, even when the null hypothesis holds exactly. The question that bears scientific interest always concerns the extent to which observed effects generalize to the population. One could argue that sometimes the authors do not mean literally what they write in these cases and that expert readers can reach the proper interpretation. Nevertheless, these expressions represent a type of miscommunication that can create ambiguity for experts and misunderstanding for lay readers. Despite much recent discourse on methodological challenges in the empirical sciences (e.g., Munafò et al., 2017), the ways in which nonsignificant findings are discussed and interpreted have remained relatively unexplored. One previous study (Hoekstra, Finch, Kiers, & Johnson, 2006) explored whether the recommendations of the fifth edition of the American Psychological Association's (2001) publication manual improved the way authors report and interpret the results of significance testing. The data showed that both before and after the publication of the new guidelines, nonsignificant effects were interpreted as claims of no effect in 60% of cases.

In this article, we report an observational study in which we investigated the prevalence of various interpretations of nonsignificant findings. We also explored the evidential value of these results using Bayes factors (e.g., Jeffreys, 1961; Kass & Raftery, 1995). Unlike NHST, Bayes factors indicate how much the data favor one hypothesis over another (Dienes, 2008). Therefore, when the necessary information was available, we computed Bayes factors for all reported nonsignificant *t*-test results in our sample. This allowed us to explore the degree to which reported nonsignificant results actually provide support for the null hypothesis.

Disclosures

Preregistration

Our data-analysis plan was uploaded to the Open Science Framework (OSF) prior to conducting the analyses. The document is available at https://osf.io/f2n7c. The statistical analyses of the link between Bayes factors and *p* values and between Bayes factors and sample sizes were not specified in the preregistration. Further minor deviations from this plan are described in the Supplemental Material available online (http://journals.sagepub.com/doi/suppl/10.1177/2515245918773742).

Data, materials, and online resources

All data and materials, as well as the R code for the analyses and figures, can be accessed at https://osf.io/f2n7c/. The Supplemental Material (http://journals.sagepub.com/doi/suppl/10.1177/2515245918773742) presents the Bayesian robustness test, describes our deviation from the original research plan, and discusses our results from exploratory analyses using the distribution of *p* values.

Reporting

We report how we determined all data exclusions and all measures in the study.

Ethical approval

No ethical approval was required for this project as we analyzed published articles without collecting new data.

Method

Sample

We selected the abstracts of every empirical research article with human participants published in 2015 in the journals *Psychological Science* (n = 150), *Psychonomic Bulletin & Review* (n = 167), and *Journal of Experimental Psychology: General* (n = 95; overall N = 412). All three are prominent journals that cover broad areas of psychological research. From this collection, we selected the articles that contained at least one negative empirical statement in their abstracts. By "negative statement," we mean that the authors explicitly stated the absence of an effect (e.g., "had no effect," "were the same") or that they referred to a nonsignificant finding (e.g., "was not significant").

For each negative statement, we screened the main text and supplement of the article to additionally record (a) the corresponding p value, (b) the type of statistical analysis, and (c) the sentence describing the results of the analysis. Additionally, when the claim was based on a t statistic (one-sample, paired-samples, or independent-samples t test), we recorded the t value and the number of participants in each experimental group.[2]

Screening procedure

The data-collection procedure was the following: One author screened the selected abstracts and judged whether they contained negative statements. If an abstract contained one or more such statements, the author extracted the corresponding additional data from the article. The selected articles were then reexamined by another author to ensure that the statements in the abstracts were based on the selected statistical tests. Next, two authors independently categorized each of the extracted claims from the abstracts using three categories and two subcategories (see Table 1 for hypothetical examples):

- The *correct-frequentist* category included statements that referred only to the fact that the analysis did not yield a significant result and did not imply that the effect is absent in the population
- The *incorrect-frequentist* category included statements indicating that the authors inferred the absence of an effect from a nonsignificant result. We differentiated two subcategories: one for statements generalized to the whole population and another for statements restricted to the current sample.
- The *Bayesian analysis* category included statements indicating that the authors used Bayes factors to quantify evidence in favor of the null hypothesis.

Disagreements regarding categorization were resolved by discussion, and the agreement of at least three authors was needed to reach a conclusion in any given case.

Calculating Bayes factors

To gauge the strength of evidence for the null hypothesis, we calculated Bayes factors, that is, the likelihood of the data under the null hypothesis (i.e., equal population means) divided by the likelihood of the data under the alternative hypothesis. Bayes factors greater than 1 indicate relative evidence for the null hypothesis, whereas Bayes factors smaller than 1 indicate relative evidence for the alternative hypothesis. As an aid for interpretation of the Bayes factors, we employed Jeffreys's (1961) classification scheme (see also Lee & Wagenmakers, 2013): Bayes factors between 1/3 and 3 are labeled *anecdotal evidence*, Bayes factors between 3 and 10 (or between 1/3 and 1/10) indicate *moderate evidence*, and Bayes factors greater than 10 or smaller than 1/10 indicate *strong evidence*.

We calculated Bayes factors only when t tests were reported. To obtain the Bayes factors that correspond to the reported t statistics and degrees of freedom, we applied the default settings of the *ttest.tstat* function of the BayesFactor R package (Morey, Rouder, & Jamil, 2015). The default settings specify the alternative hypotheses by assigning effect size a two-tailed Cauchy

Table 1. Hypothetical Examples for the Categories of Claims Concerning Nonsignificant Findings

Category	Example
Correct-frequentist	"The analysis did not show a significant effect of the intervention."
Incorrect-frequentist: whole population	"The results establish that the intervention has no effect on the dependent variable."
Incorrect-frequentist: current sample	"There was no difference between the participants in the intervention group and the control group."
Bayesian analysis	"The Bayes factor favored the null hypothesis over the alternative hypothesis."

distribution with medium scale (i.e., $r = \sqrt{2}/2$). This *default JZS prior* (Rouder, Speckman, Sun, Morey, & Iverson, 2009) constitutes one of several proposed methods to specify the predictions of the alternative hypothesis. As we detail later, we repeated our Bayes factor reanalysis using two alternative prior distributions in order to explore the robustness of the results.

Results

Planned analyses

Screening. We found at least one negative statement in 132 of the 412 screened abstracts (*Psychological Science:* $n = 39$; *Psychonomic Bulletin & Review:* $n = 58$; *Journal of Experimental Psychology: General:* $n = 35$). These 132 abstracts contained 137 negative statements (*Psychological Science:* $n = 39$; *Psychonomic Bulletin & Review:* $n = 61$; *Journal of Experimental Psychology: General:* $n = 37$). We linked these statements to 175 statistical tests from the articles, and we collected 122 reported p values from these tests (*Psychological Science:* $n = 26$; *Psychonomic Bulletin & Review:* $n = 46$; *Journal of Experimental Psychology: General:* $n = 50$). The number of reported p values is substantially smaller than the number of tests because some tests used nonfrequentist statistics (e.g., Bayes factors) and in several cases, the p value was not reported (e.g., for nonsignificant regression slopes or analyses of variance) and could not be retrieved from the authors.[3]

Categories of statements. We found that 72% ($n = 98$) of the negative statements misinterpreted the nonsignificant result; 23% ($n = 32$) fell in the "incorrect-frequentist: whole population" subcategory, and 48% ($n = 66$) fell in the "incorrect-frequentist: current sample" subcategory. Only 18% ($n = 25$) of the statements were categorized as correct frequentist reporting. The least common category was "Bayesian analysis," which included only 10% ($n = 14$) of the statements. Table 2 reports the frequencies of the different categories of negative claims broken down by journal.

Bayesian analyses. From the 175 statistical tests that we collected from the articles, we identified 67 t tests and were able to acquire the necessary information for Bayesian analyses of 63 tests. We calculated Bayes factors (BF_{01}—evidence in favor of the null hypothesis) with a medium-scale ($r = \sqrt{2}/2$) Cauchy prior under the alternative hypothesis. The 63 t tests yielded 16 anecdotal (25%), 45 moderate (71%), and 2 strong (3%) BF_{01}s, all of them in favor of the null hypothesis. Both of the strong BF_{01}s were obtained in studies with sample sizes of more than 300 participants (see Exploratory Bayesian Analyses for a more thorough description of the link between sample size and BF_{01}s).

Robustness test. These results were obtained for a specific prior distribution (i.e., a two-tailed medium-scale Cauchy distribution on the standardized effect size). To probe the robustness of the results, we calculated the BF_{01}s of the 63 t tests using normal priors (Dienes, 2014) and informed priors (Gronau, Ly, & Wagenmakers, 2017; see the Supplemental Materials for a detailed description). Figure 1 shows the BF_{01}s, ordered by their size, as calculated with each of the three priors. The figure also indicates the percentages of the BF_{01}s in the different evidence categories. With the default prior, 74.6% ($n = 47$) of the BF_{01}s were greater than 3 (providing at least moderate evidence for the null), whereas with the informed prior, only 44.4% ($n = 28$) of the BF_{01}s provided this level of support for the null. BF_{01}s computed with the normal prior showed even weaker evidential support for the null, as only 25.4% ($n = 16$) of them exceeded a value of 3. Applying the informed rather than the default prior changed the evidential category of the BF_{01}s in 20 cases (31.7%), and application of the normal rather than the default prior resulted in 33 (52.4%) changes in the evidential category. However, as is apparent from Figure 1, the differences between the values of the BF_{01}s calculated with the different models were in most cases not substantial. The large number of differences in evidence categorizations is due to the fact that the majority of the BF_{01}s were scattered around the category thresholds.

Table 2. Frequencies of the Negative Statements Broken Down by Category and Journal

Category	*Psychological Science*	*Psychonomic Bulletin & Review*	*Journal of Experimental Psychology: General*	Total
Correct-frequentist	4	9	12	25
Incorrect-frequentist: whole population	7	15	10	32
Incorrect-frequentist: current sample	25	29	12	66
Bayesian analysis	3	8	3	14
Total	39	61	37	137

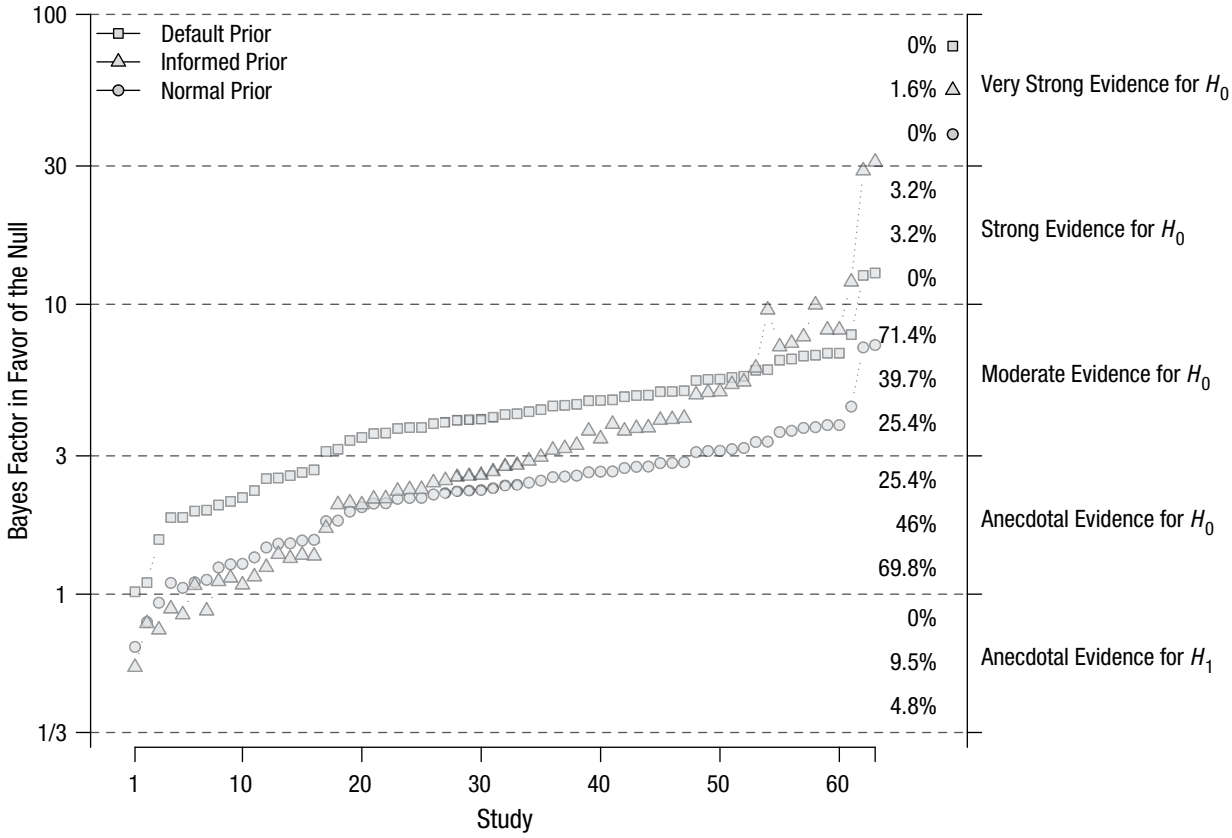

Fig. 1. Bayes factors in favor of the null hypothesis (H_0) for each of the 63 nonsignificant *t* tests reported in the selected literature. For each *t* test, Bayes factors were calculated with default, informed, and normal prior specifications of the alternative hypothesis (H_1). Note that the scaling of the *y*-axis has been log-e transformed to facilitate visualization of the relationships between the Bayes factors calculated with different prior specifications. The labels on the right-hand side of the *y*-axis represent Jeffreys's (1961) scheme for classifying the strength of evidence. To the left of each label, the numbers indicate the percentage of all results falling in the indicated category when the Bayes factors were calculated using default, informed, and normal prior specifications, respectively (from top to bottom). This figure has been reformatted from the original, which was published under a CC-By Attribution International License and is available at https://doi.org/10.6084/m9.figshare.5721076.v1.

Exploratory Bayesian analyses

To explore the extent to which the *p* values and the corresponding BF_{01}s were associated, we plotted the reported *p* values[4] against the BF_{01}s calculated with the default prior (see Fig. 2) and conducted Bayesian parameter estimation by computing Kendall's τ and its 95% credible interval (CI; see Box 1 for additional analytic details). This correlation analysis revealed that the relationship between the *p* values and the BF_{01}s was moderate and that the true value of the correlation likely fell between .20 and .50 (τ = .38, 95% CI = [.20, .52]). Figure 2 shows that this moderate relation was driven primarily by the correlation between the low *p* values (smaller than .3) and the BF_{01}s, and that the values of the BF_{01}s leveled off for *p* values higher than .3. The figure also shows that high *p* values do not guarantee strong evidence for the null hypothesis.

Next, we investigated the relationship between sample size and BF_{01}. It is apparent from Figure 3 that the majority of the BF_{01}s providing anecdotal evidence in favor of the null hypothesis (13 cases, 81.25% of all anecdotal BFs) were obtained in studies with small sample sizes ($n < 35$). In contrast, 48% (12 cases) of the small samples produced moderate evidence in favor of the null hypothesis. Strong evidence was reached only in studies with large samples ($n > 300$), and all of the large-sample studies provided at least moderate evidence in favor of the null hypothesis. To estimate the strength of the association between sample size and BF_{01}, we calculated the correlation coefficient and its 95% CI. We found a positive correlation (τ = .45, 95% CI = [.26, .59]).

Discussion

The goal of this study was twofold: to explore how psychology researchers interpret and communicate nonsignificant results and to assess how much these results truly constitute evidence in favor of the null

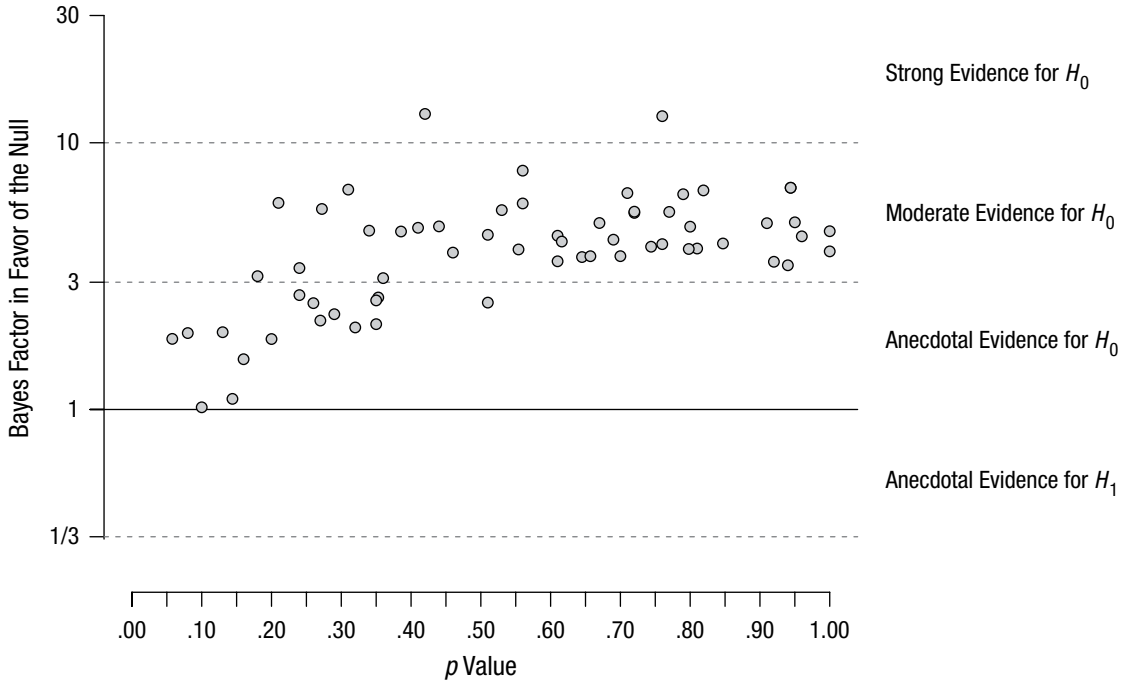

Fig. 2. Scatterplot showing the relationship between the *p* values from the selected studies and the corresponding default BF_{01}s (τ = .38, 95% credible interval = [.20, .52]). Note that the scaling of the *y*-axis has been log-e transformed to facilitate visualization of the relationship. Plotted points above the solid black line indicate evidence for the null hypothesis. This figure has been reformatted from the original, which was published under a CC-By Attribution International License and is available at https://doi.org/10.6084/m9.figshare.5721076.v1.

hypothesis. Toward this end, we collected all the negative statements from the abstracts of three domain-general psychology journals, and we extracted and reanalyzed the statistics corresponding to those statements.

The analysis of the negative statements in the abstracts demonstrates that there are several ways in which researchers interpret nonsignificant results. Only 28% (n = 39) of these statements were in agreement with the logic of the employed statistical methods (frequentist: 18%, n = 25; Bayesian: 10%, n = 14). Among the incorrect inferences, the smaller fraction of the statements (23%, n = 32) indicated that there was no effect in the population. The most prevalent interpretation of nonsignificant results, however, was incorrect and limited to the observed sample (48%, n = 66). Although it is possible that the words the researchers used to describe their results did not reflect what they meant to say, awareness of this habit must be raised because interpreting the results of an inferential test with respect to the observed sample is not meaningful.

In an exploratory analysis reported in the Supplemental Material, we compared the extracted statistical results (i.e., those corresponding to negative statements in abstracts) with all the reported nonsignificant statistical results from the same year in the same three journals. This analysis suggests that researchers are less likely to build an argument on a nonsignificant result if the corresponding *p* value is small than if it is large.

These observations underscore the apparent confusion and uncertainty regarding the interpretation of

Box 1. Bayesian Parameter Estimation

For the two exploratory correlation analyses reported in this section, we decided to conduct Bayesian parameter estimation instead of hypothesis testing. Therefore, we report the correlation coefficients (Kendall's τs) with their 95% credible intervals (CIs). The investigated associations were nonlinear; thus, we opted to compute Kendall's τ to estimate the population effect sizes (e.g., Kendall & Gibbons, 1990). To calculate Kendall's τ, we used the *KendallTauB* function from the DescTools R package (Signorell, 2017). We passed on the τ value and the sample size to compute the 95% CIs with the *credibleIntervalKendallTau* function created by van Doorn, Ly, Marsman, and Wagenmakers (2016). We employed the two-tailed default prior distribution of τ, which is a nonuniform distribution on τ constructed from a uniform distribution on the Pearson's ρ (parametric yoking; van Doorn et al., 2016).

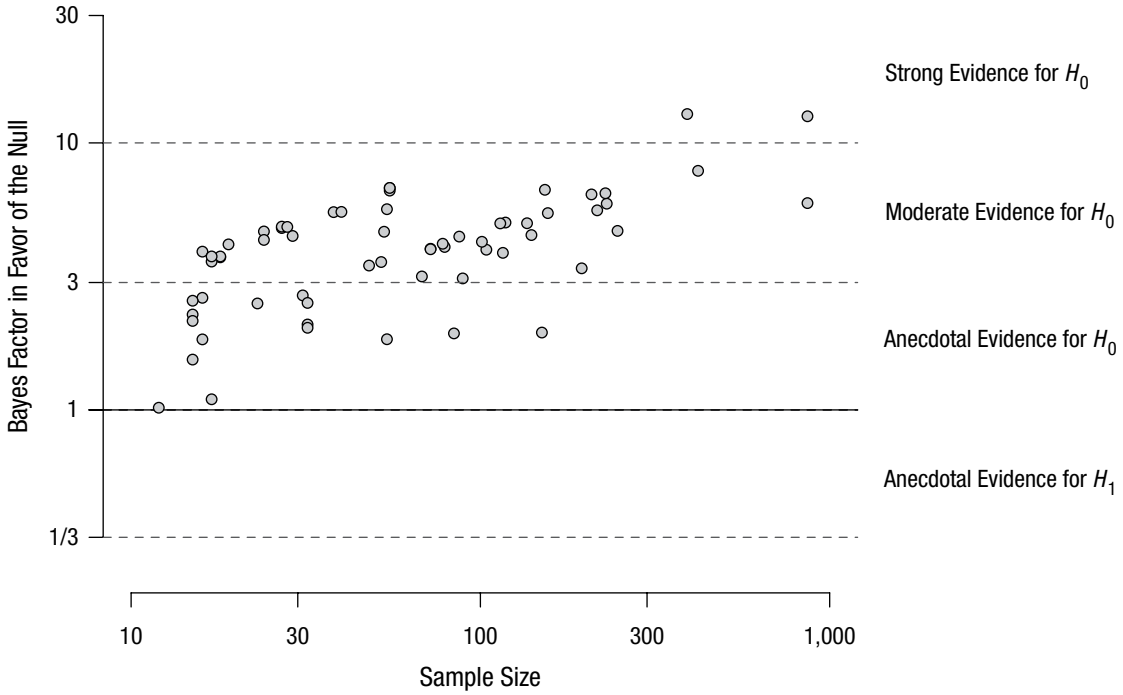

Fig. 3. Scatterplot showing the relationship between the sample sizes of the selected studies and the corresponding BF_{01}s (τ = .45, 95% credible interval = [.26, .59]). Note that the scaling of both axes has been log-e transformed. This figure has been reformatted from the original, which was published under a CC-By Attribution International License and is available at https://doi.org/10.6084/m9.figshare.5721076.v1.

nonsignificant results, and they also reflect that the field has no generally applied strategy for discussing nonsignificant findings. The apparent confusion and uncertainty in research practice possibly originate from the fact that although researchers are motivated to discuss all of their findings, the NHST framework is not designed to be informative about negative results (Fisher, 1935). As Fisher (1935) wrote:

> It should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (p. 19)

This limitation has resulted not just in a great number of uncommunicated negative results (Franco et al., 2014), but also, as this study shows, in unwarranted interpretation of negative findings.

To assess the extent to which the reported nonsignificant findings in our sample constitute the absence of evidence (i.e., nondiagnostic results produced by low power) or evidence of absence (i.e., support for the null hypothesis), we conducted a Bayes factor reanalysis. The interpretation of Bayes factors is always conditional on the level of support for the hypotheses expected from the data (Aczel, Palfi, & Szaszi, 2017). As is apparent in Figure 2, almost all the BF_{01}s were smaller than 10, and a great proportion of them were under 3. Although there are different traditions for characterizing the strength of evidence indicated by Bayes factors (Schönbrodt, 2015), values lower than 3 are most often interpreted as anecdotal, and values lower than 10 are generally not considered strong evidence (Lee & Wagenmakers, 2013). Note that when the null hypothesis and the alternative hypothesis are deemed equally likely a priori, a Bayes factor of 3 raises the model probability for the null hypothesis from 50% to 75% (leaving a full 25% probability for the alternative hypothesis), and a Bayes factor of 10 raises the model probability from 50% to 91% (leaving 9% probability for the alternative hypothesis).

The result of our reanalysis, and its robustness to alternative prior specifications of the alternative hypothesis, suggests that the nonsignificant findings that were elevated to the abstracts of the selected studies provide at best only moderate evidence for the authors' negative claim. In a considerable number of cases, the nonsignificant findings presented in the abstracts carry evidence that is not worth more than a bare mention (Etz & Vandekerckhove, 2017; Jeffreys, 1961). This weakness may be partly due to the typically low sample sizes in psychology (see, e.g., Aczel, Palfi, Szaszi, Szollosi, & Dienes, 2015; Kekecs et al., 2016). Hoekstra, Monden, van Ravenzwaaij, and Wagenmakers (2018), reanalyzing

nonsignificant results in medicine, found much stronger evidence for the null hypothesis with samples two or three magnitudes larger than those in our analysis. Our finding of a moderate link between sample size and Bayes factor further corroborates this explanation.

Taken together, our results extend the list of reasons why current research practice in psychological science needs to be reconsidered. It is a long-known problem that positive results are more attractive (Giner-Sorolla, 2012) and more likely to be published (Franco et al., 2014; Rosenthal, 1979) than negative results. This publication bias is often blamed for promulgating misleading and nonreplicable findings and for resulting in a loss of immense resources (Lilienfeld & Waldman, 2017). Here, we have shown that even when these negative findings are reported, they are often miscommunicated or lack sufficient evidential support. In fact, the situation has not improved since Hoekstra et al. (2006) observed that 61% of the psychology articles published between 2002 and 2004 claimed no effect or a negligible effect purely on the basis of statistically nonsignificant results. We suggest that negative results in science carry a "curse" that is due not only to their lack of attraction, but also to the problematic status of negative results within the NHST tradition, as well as to the chronic underestimation of required sample sizes in psychological experiments.

We note that our sample for the Bayesian reanalysis was constrained to *t* tests in articles published in three journals in 2015. Nevertheless, we would not expect to obtain a substantially different pattern of results with a more comprehensive sample given that a recent Bayesian reanalysis of more than 300,000 published significant *t*-, *F*- and *r*-test results indicated that the strength of evidence is comparable among the different statistical tests in psychological studies (Aczel et al., 2017). The generalizability of any Bayesian analysis depends on the predictions of the tested hypotheses, which is determined by their prior distributions. We examined the robustness of our conclusions with a range of different prior distributions, and each time we obtained the same pattern of results.

Transparency in conducting and communicating research is of primary importance for improving the field. However, the field may also benefit from adopting a more inclusive statistical approach. For instance, the proponents of Bayes factors argue that Bayesian analysis could help alleviate several of the current challenges. Bayes factors can be interpreted as evidence not just against, but also for, the null hypothesis. In addition, they are insensitive to stopping rules, allowing the experimenter to stop data collection whenever the evidence for one of the hypotheses is sufficiently compelling (Dienes, 2016; Rouder, 2014; but see de Heide & Grünwald, 2017). The Bayes factor is not the only tool for testing the absence of an effect or demonstrating that an effect is too small to be practically relevant. For instance, parameter estimation with confidence intervals (e.g., Cumming, 2014) can be informative about the size of an effect, and equivalence testing (Lakens, 2017), a frequentist procedure that is conceptually similar to analysis of the Bayesian region of practical equivalence (ROPE; e.g., Kruschke, 2014), provides a way to accept the null hypothesis if a region of negligible effect sizes can be determined. Nonetheless, these alternative methods cannot be applied to test a point null hypothesis, which was the primary focus of the current study.

It has long been recognized that psychological experiments are often underpowered (Cohen, 1990). The statistical power of a typical two-group between-subjects design is estimated to be less than 35% (Bakker, van Dijk, & Wicherts, 2012), and power analysis is reported for only 3% of psychological studies in general. Although these issues might be traced back to some inappropriate rules of thumb existing among research psychologists (Bakker, Hartgerink, Wicherts, & van der Maas, 2016), our results provide further evidence that without a substantial increase in statistical power, psychologists' data can provide only weak evidence in favor of the null hypothesis.

Conclusion

Our findings reveal that nonsignificant results are often misinterpreted in the psychology literature. Moreover, our Bayesian reanalyses reveal that most nonsignificant findings reported in the abstracts in this literature provide only limited evidence for the null hypothesis. These observations suggest that nonsignificant findings, as traditionally reported, can easily mislead the reader. Specific statistical training, a more skeptical mind-set, and an extension of the standard statistical toolbox are possible remedies to promote more adequate communication and more appropriate assessment of negative results.

Action Editor

Alex O. Holcombe served as action editor for this article.

Author Contributions

B. Aczel, B. Palfi, A. Szollosi, B. Szaszi, and E.-J. Wagenmakers designed the study and wrote the manuscript. M. Kovacs, P. Szecsi, and M. Zrubka contributed to the data collection and methodology. B. Palfi, M. Kovacs, Q. F. Gronau, and D. van den Bergh contributed to the analysis and visual presentation of the results. All the authors reviewed and approved the final version of the submitted manuscript.

ORCID iDs

Aba Szollosi https://orcid.org/0000-0003-3457-542X
Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1034

Acknowledgments

We thank Maarten Marsman for his help with the code and Melissa Wood for her comments on earlier versions of the manuscript. B. Palfi is grateful to the Dr Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

A. Szollosi was supported by the "Nemzet Fiatal Tehetségeiért" Scholarship (NTP-NFTÖ-16-1184), and E.-J. Wagenmakers was supported by a Vici grant from the Netherlands Organisation for Scientific Research (NWO, 016.Vici.170.083).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918773742

Open Practices

All data and materials, as well as the R code for the analyses and figures, have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/f2n7c/. The analysis plans can be accessed at https://osf.io/f2n7c/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918773742. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Throughout this article, we use *significant* and *nonsignificant* to refer to statistical, not theoretical, significance.
2. When the data were not available, we requested them from the corresponding author via e-mail. When group sizes were not provided for independent-samples *t* tests, we took half of the total number of participants as the sample size of each group. When the exact *p* value of a *t* test was not reported, we calculated it from the *t* and degrees of freedom, if these values were available.
3. We contacted 19 authors in total; 4 did not reply, 10 provided the required information, and 5 did not provide the required information.
4. Note that 4 of the 63 *p* values were obtained from one-tailed tests. As the focus of our interest was how researchers interpret nonsignificant *p* values, we did not transform these values to correspond to the results of two-tailed tests.

References

Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE*, *12*(8), Article e0182651. doi:10.1371/journal.pone.0182651

Aczel, B., Palfi, B., Szaszi, B., Szollosi, A., & Dienes, Z. (2015). Commentary: Unlearning implicit social biases during sleep. *Frontiers in Psychology, 6*, Article 1428. doi:10.3389/fpsyg.2015.01428

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*, 1069–1077.

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

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

de Heide, R., & Grünwald, P. D. (2017). *Why optional stopping is a problem for Bayesians*. Retrieved from https://arxiv.org/pdf/1708.08278.pdf

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Basingstoke, England: Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article 781. doi:10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:10.1016/j.jmp.2015.10.003

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi:10.3758/s13423-017-1262-3

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*, 555–561. doi:10.1177/1745691612459059

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505. doi:10.1126/science.1255484

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571. doi:10.1177/1745691612457576

Goodman, S. (2008). A dirty dozen: Twelve *P*-value misconceptions. *Seminars in Hematology*, *45*, 135–140. doi:10.1053/j.seminhematol.2008.04.003

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. doi:10.1037/h0076157

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). *Informed Bayesian t-tests*. Retrieved from https://arxiv.org/pdf/1704.02479.pdf

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than. 05 in psychology: What is going on? *PeerJ*, *4*, Article e1935. doi:10.7717/peerj.1935

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology*, *3*, Article 9. doi:10.1525/collabra.71

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review*, *13*, 1033–1037.

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). *Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. PLOS ONE*, *13*(4), Article e0195474. doi:10.1371/journal.pone.0195474

Jeffreys, H. (1961). *The theory of probability*. Oxford, England: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. doi:10.1080/01621459.1995.10476572

Kekecs, Z., Szollosi, A., Palfi, B., Szaszi, B., Kovacs, K. J., Dienes, Z., & Aczel, B. (2016). Commentary: Oxytocin-gaze positive loop and the coevolution of human–dog bonds. *Frontiers in Neuroscience*, *10*, Article 155. doi:10.3389/fnins.2016.00155

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London, England: Edward Arnold.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological & Personality Science*, *8*, 355–362.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. Chichester, England: John Wiley & Sons.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12-2) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. doi:10.1038/s41562-016-0021

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*, 1036–1042. doi:10.1177/0956797616645672

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

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. doi:10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:10.3758/PBR.16.2.225

Schönbrodt, F. (2015, April 17). Grades of evidence – a cheat sheet [Web log post]. Retrieved from http://www.nicebread.de/grades-of-evidence-a-cheat-sheet/

Signorell, A. (2017). DescTools: Tools for descriptive statistics (Version 0.99.22) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/DescTools/index.html

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2016). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*. Advance online publication. doi:10.1080/00031305.2016.1264998

Wilkinson, L., & the Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi:10.1037/0003-066X.54.8.594