Interpreting Non-significant Results: A Quantitative Analysis Based on 500 Empirical Studies

Authors: Wang Jun, Song Qiongya, Xu Yuepei, Jia Binbin, Lu Chunlei, Chen Xi, Dai Zixu, Huang Zhiyue, Li Zhenjiang, Lin Jingxi, Luo Wanying, Sainan Shi, Zhang Yingying, Zang Yufeng, Zuo Xinian, Hu Chuanpeng, Hu Chuanpeng

Date: 2020-10-17T00:00:00+00:00

Abstract

Nonsignificant results (e.g., $p > 0.05$) are very common in psychological research and can be easily misinterpreted as evidence for accepting the null hypothesis, potentially leading to erroneous inferences in group-matching studies or overlooking true effects masked by nonsignificant results from small samples. However, there is currently no empirical research domestically investigating the prevalence of nonsignificant results and their interpretation. This study examined 500 Chinese-language empirical psychological research articles, statistically analyzing the frequency of negative statements related to nonsignificant results in their abstracts, assessing and tallying the accuracy of inferences based on these negative statements, and re-evaluating studies with t-values in their nonsignificant results using Bayes factors. The results showed that 36% of abstracts mentioned nonsignificant results, containing a total of 236 negative statements. Among these, 41% of the negative statements exhibited biased interpretation of nonsignificant results (e.g., interpreting them as supporting the null hypothesis). Bayes factor analysis of studies containing t-values revealed that only 5.1% of nonsignificant results could provide strong evidence supporting the null hypothesis ($BF01 > 10$). Compared with previous survey results from international psychology journals (30% of abstracts contained negative statements; 70% of negative statements misinterpreted nonsignificant results), Chinese psychology journals showed both a higher proportion of reporting nonsignificant results and a higher accuracy rate in interpreting them. However, domestic researchers still need to further enhance their understanding of nonsignificant results and promote statistical methods suitable for evaluating them.

Full Text

Interpreting Nonsignificant Results: A Quantitative Analysis Based on 500 Empirical Studies

Jun Wang[1], Qiongya Song[1], Yuepei Xu[2,3], Binbin Jia , Chunlei Lu , Xi Chen , Zixu Dai , Zhiyue Huang , Zhenjiang Li , Jingxi Lin[1] , Wanying Luo[11], Sainan Shi[12], Yingying Zhang[13], Yufeng Zang[1] , Xinian Zuo[2], Chuanpeng Hu[1]

[1]Department of Psychology, Sun Yat-sen University, Guangzhou, 510006, China
[2]Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
[3]Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China
School of Psychology, Shanghai University of Sport, Shanghai, 200438, China
College of Teacher Education, Zhejiang Normal University, Jinhua, 321000, China
School of Psychology, Shanghai University of Finance and Economics, Shanghai, 200122, China
School of Psychology, South China Normal University, Guangzhou, 510631, China
Tisch School of the Arts, New York University, New York, 11201, USA
School of Education, Soochow University, Suzhou, 215123, China
[1] Institute of Education Science, Heilongjiang University, Harbin, 150080, China
[11]School of Psychological and Cognitive Sciences, Peking University, Beijing, 100871, China
[12]School of Psychology and Cognitive Science, East China Normal University, Shanghai, 200063, China
[13]Faculty of Psychology, Southwest University, Chongqing, 400715, China
[1] Center for Cognition and Brain Disorders, Hangzhou Normal University, Hangzhou, 311121, China
[1] Leibniz Institute for Resilience Research, 55131 Mainz, Germany

Author Contributions: This study adopted the CREDIT taxonomy for research contributions (https://casrai.org/credit/) and defined 14 contribution domains. Jun Wang contributed to data curation, formal analysis, investigation, methodology, resources, writing—original draft, and writing—review and editing. Qiongya Song, Yuepei Xu, and Chunlei Lu contributed to investigation, methodology, resources, and writing—review and editing. Binbin Jia contributed to investigation, methodology, resources, visualization, validation, and writing—review and editing. Xi Chen, Zixu Dai, Zhiyue Huang, Zhenjiang Li, Jingxi Lin, Wanying Luo, Sainan Shi, and Yingying Zhang contributed to investigation. Yufeng Zang and Xinian Zuo contributed to supervision and writing —review and editing. Chuanpeng Hu contributed to conceptualization, investigation, methodology, project administration, supervision, and writing—review and editing. Individual author contributions are illustrated in the figure below.

Corresponding Author: Chuanpeng Hu, Email: hcp4715@hotmail.com

Abstract

Nonsignificant results (e.g., p > 0.05) are common in psychological research but are often misinterpreted as evidence supporting the null hypothesis, potentially leading to erroneous inferences in group-matching studies or causing researchers to overlook true effects masked by small-sample nonsignificant results. However, no empirical studies in China have investigated the prevalence of nonsignificant results and how they are interpreted. This study examined 500 Chinese psychological empirical research articles, quantifying the frequency of negative statements related to nonsignificant results in abstracts, evaluating the accuracy of inferences based on these statements, and using Bayes factors to re-evaluate studies reporting nonsignificant results with t-values. Results showed that 36% of abstracts mentioned nonsignificant results, containing a total of 236 negative statements. Among these, 41% misinterpreted the nonsignificant results (e.g., interpreting them as supporting the null hypothesis). Bayesian factor analysis of studies reporting t-values revealed that only 5.1% of nonsignificant results provided strong evidence supporting the null hypothesis (BF > 10). Compared with previous surveys of international psychology journals (30% of abstracts contained negative statements; 70% misinterpreted nonsignificant results), Chinese psychology journals showed both a higher proportion of reported nonsignificant results and higher accuracy in their interpretation. Nevertheless, Chinese researchers need to strengthen their understanding of nonsignificant results and promote statistical methods appropriate for evaluating them.

Keywords: Nonsignificant results; Null hypothesis significance testing; Bayes factors; Meta-research

1. Introduction

Appropriate statistical inference methods are essential for drawing correct conclusions from data. Currently, null hypothesis significance testing (NHST) dominates scientific practice (American Psychological Association, 2010; Wasserstein & Lazar, 2016). Within this framework, researchers typically make a binary decision about whether to reject the null hypothesis based on the p-value. Specifically, when the p-value falls below a predetermined threshold (typically set at 0.05), researchers can reject the null hypothesis and accept the alternative hypothesis; when the p-value exceeds this threshold, researchers fail to reject the null hypothesis. However, failure to reject the null hypothesis entails two possibilities: first, the data support the null hypothesis, meaning the effect is absent (evidence of absence); second, the study lacks sufficient statistical power to detect a true effect (Dienes, 2014, 2016), meaning there is absence of evidence.

Researchers have long recognized the limitations of NHST (Amrhein et al., 2019; Edwards et al., 1963; Gigerenzer et al., 2004; Meehl, 1967; Miller, 2011; Nickerson, 2000; Ziliak & McCloskey, 2008). On one hand, the dichotomous decision-

making mindset inherent in NHST has contributed to the neglect and even stigmatization of nonsignificant results, leading to publication bias in the scholarly literature (Sun et al., 2012). Fanelli (2012) analyzed publications across disciplines and found that the proportion of positive/significant results exceeded that of negative/nonsignificant results in all fields, with psychology showing a particularly high proportion of positive results at over 95%. Such publication bias may lead to erroneous estimates of true effects (Algermissen & Mehler, 2018; Schäfer & Schwarz, 2019), contributing to the reproducibility crisis in psychology (Baker, 2016; Ioannidis, 2005; Klein et al., 2014; Open Science Collaboration, 2015; Hu et al., 2016). On the other hand, researchers sometimes misinterpret nonsignificant results. Although $p > 0.05$ cannot distinguish between "data support the null hypothesis" and "data are insufficient to support or reject the null hypothesis," researchers often conflate these situations in their conclusions, erroneously treating $p > 0.05$ as evidence supporting the null hypothesis, thereby compromising the credibility of their findings (Greenland et al., 2016; X. Lyu et al., 2020; Z. Lyu et al., 2018; Hu et al., 2016; Luo, 2017). Lyu et al. (2020) found that 53% of researchers mistakenly believed that $p > 0.05$ indicates data support for the null hypothesis.

Such misinterpretation of nonsignificant results can have two serious consequences. First, erroneously accepting the null hypothesis can affect subsequent inferences about intervention effects. In clinical trials, researchers often use chi-square tests or independent samples t-tests to analyze differences between experimental and control groups on confounding variables (e.g., gender, age, education level). When a t-test yields $p > 0.05$ (e.g., 0.06), researchers may conclude that the groups do not differ on that variable and subsequently disregard its potential influence when analyzing intervention effects, overlooking serious confounding. Second, misinterpretation can lead to neglect of negative results. Researchers may lack sufficient power to detect existing effects due to small sample sizes, resulting in nonsignificant findings (Button et al., 2013; Chen et al., 2018). If they misinterpret these nonsignificant results as indicating no effect, they may miss potentially important effects (Fiedler et al., 2012). For example, a multi-center meta-analysis showed that although the left putamen was the most abnormal brain region in Parkinson's patients at the meta-analytic level, individual center results—due to low power—found significant putamen abnormalities after multiple comparison correction in only 2 out of 18 centers (Jia et al., 2018).

Although discussions about nonsignificant results within the NHST framework have increased (Lü, 2014; Zhong, 2016), most are theoretical or methodological, lacking empirical investigations into the prevalence and interpretation of nonsignificant results in Chinese psychological literature. Aczel et al. (2018) reviewed 412 empirical articles published in 2015 in *Psychonomic Bulletin & Review*, *Journal of Experimental Psychology: General*, and *Psychological Science*, finding that nearly one-third of abstracts contained negative statements (where researchers explicitly stated that an effect was absent or mentioned nonsignificant results), with 72% of these articles misinterpreting nonsignificant

results. Do similar misinterpretations exist in authoritative Chinese psychology journals?

Moreover, researchers sometimes genuinely need to demonstrate null effects or that the null hypothesis is true. As noted above, for group-matching in between-subjects designs, researchers need to ensure equivalence between experimental and control groups on certain attributes (e.g., age, gender). In such cases, researchers must provide evidence for the null hypothesis that "there are no other differences,"and misinterpreting p-values leads to erroneous statistical inferences. Researchers may also need to test competing theories, using experimental data to show that a difference predicted by one theory does not exist—that is, supporting the null hypothesis. In other words, in certain contexts, demonstrating that "the null hypothesis is true" is the intended goal, serving to reject or falsify a research hypothesis and propose alternative hypotheses, thereby advancing scientific theory.

Since NHST cannot provide support for the null hypothesis, using $p > 0.05$ as evidence for it is actually incorrect (Chuard et al., 2019). Researchers therefore need to introduce appropriate statistical methods to assess the degree to which data support the null hypothesis, such as Bayes factors (BFs) (Wagenmakers et al., 2018; Wagenmakers et al., 2011; Hu et al., 2018). Aczel et al. (2018) recalculated Bayes factors for data with nonsignificant t-test results to assess the degree of support for the null hypothesis, finding that only 3% of these nonsignificant t-tests provided strong evidence supporting the null hypothesis ($BF > 10$), while 71% provided only moderate evidence ($10 > BF > 3$). These results suggest that without appropriate statistical methods, researchers may overlook an important issue: data yielding nonsignificant results often cannot provide sufficiently strong evidence for the null hypothesis. Whether this phenomenon exists in published articles from Chinese psychology core journals is also worth exploring, as understanding it can help Chinese psychology researchers recognize that misinterpreting nonsignificant results yields erroneous evidence supporting the null hypothesis.

To obtain empirical data on the current state of nonsignificant result interpretation in Chinese psychology research, this study—following Aczel et al. (2018)—surveyed empirical research articles published in 2017 and 2018 in five Chinese psychology core journals (*Acta Psychologica Sinica*, *Psychological Science*, *Chinese Journal of Clinical Psychology*, *Psychological Development and Education*, and *Studies of Psychology and Behavior*). Specifically, we analyzed the reporting and misinterpretation rates of nonsignificant results in a random sample of 500 papers, recalculated Bayes factors to assess whether data with nonsignificant results could actually support the null hypothesis and to what degree, and compared the status of nonsignificant result interpretation between Chinese core journals and international journals. This article aims to raise researcher awareness of the prevalence of misinterpreting nonsignificant results, encouraging more cautious and rigorous statistical inference to avoid misinterpretation.

2. Methods

2.1 Article Sampling

This study selected five Chinese psychology core journals whose full texts were freely accessible: *Acta Psychologica Sinica*, *Psychological Science*, *Chinese Journal of Clinical Psychology*, *Psychological Development and Education*, and *Studies of Psychology and Behavior*. These journals cover diverse areas of psychological research, including cognitive, developmental, social, and clinical psychology. We compiled all empirical research articles published in these five journals during 2017–2018 (i.e., articles containing data analysis, excluding reviews, meta-analyses, or commentaries), recording the title, publication date, volume, and page numbers for each article and assigning each a unique ID. For example, the second article in *Acta Psychologica Sinica* was coded as 1002—where "1" represents the journal ID (different journals have different IDs) and "002" indicates the article' s order within the journal. Detailed coding rules are available at https://osf.io/mf42q/. Finally, we randomly sampled empirical articles from each journal proportionally to their publication volumes. The numbers of empirical articles were 246, 299, 379, 162, and 213, respectively, totaling 1,299 articles, corresponding to proportions of 18.94%, 23.02%, 29.18%, 12.47%, and 16.40%. Therefore, the numbers of randomly selected articles were: *Acta Psychologica Sinica* (95), *Psychological Science* (115), *Chinese Journal of Clinical Psychology* (146), *Psychological Development and Education* (62), and *Studies of Psychology and Behavior* (82).

The code for random sampling is available at https://osf.io/7my4g/.

2.2 Article Coding

The coding process consisted of three steps: initial coding, coding verification, and classification coding (Figure 1 [Figure 1: see original paper]). In the initial coding phase, the 500 selected articles were randomly divided into 13 batches and assigned to 13 coders. The coding procedure was as follows: coders read each article' s abstract to determine whether it contained at least one negative statement. A "negative statement" was defined as a statement in which researchers explicitly claimed that an effect was absent (e.g., "there was no difference between the intervention and control groups") or mentioned a non-significant result (e.g., "no evidence supported a significant difference between the intervention and control groups"). If the abstract contained no negative statements, coders only extracted basic article information, including article ID, citation, article link, and article type. If the abstract contained at least one negative statement, coders additionally extracted the negative statement itself and the corresponding statistical test information from the main text. This statistical information primarily included the test method; for t-tests (including one-sample, paired-samples, and independent-samples t-tests), coders also extracted t-values, p-values, and sample sizes for subsequent Bayes factor calculation.

To ensure coding accuracy, after initial coding was completed, articles were reassigned for verification. Detailed coding templates and procedures are available in the supplementary materials (https://osf.io/a39hb/).

Figure 1. Literature coding and data extraction workflow

After extracting negative statements and corresponding statistical results, six coders independently classified the negative statements, then discussed discrepancies to reach final classification. Specific categories and criteria are shown in Table 1 . To assess inter-rater agreement among the six coders, we calculated Fleiss' kappa (Fleiss, 1971) using the irr R package developed by Gamer et al. (2019) (function kappam.fleiss), which is appropriate for categorical variables with more than two raters.

Table 1. Specific categories and classification criteria for negative statements

Category	Classification Criteria
Correct frequentist interpretation	Interpreting nonsignificant results according to NHST logic, i.e., only stating that there is no evidence supporting the alternative hypothesis. Example: "The results show no evidence supporting a significant difference between the intervention and control groups."
Incorrect frequentist interpretation–Generalization to population	Interpreting nonsignificant results as supporting the null hypothesis at the population level from which the sample was drawn. Example: "The results fail to reject the null hypothesis or fail to support the alternative hypothesis."
Incorrect frequentist interpretation–Based on current sample	Interpreting nonsignificant results as supporting the null hypothesis in the current sample. Example: "The results indicate no difference between the intervention and control groups."
Bayes factor-based interpretation	Using Bayes factors to support the null hypothesis over the alternative hypothesis. BF > 10 indicates strong evidence for the null hypothesis.
Difficult to judge	Due to linguistic ambiguity, the category of the negative statement cannot be clearly determined. Example: "Except for fear emotions, the greater the intensity of basic expressions, the better participants' recognition of the expressions."

2.3 Bayes Factor Analysis

To re-evaluate the degree to which data from studies using t-tests (one-sample, paired-samples, or independent-samples t-tests) supported the null hypothesis,

we calculated Bayes factors based on reported statistical parameters (sample size and t-values) (Ly et al., 2018). Bayes factors compare the relative extent to which data support the alternative hypothesis (H) versus the null hypothesis (H) (Wagenmakers et al., 2018), using the formula:

$$BF_{01} = \frac{P(Data|H_0)}{P(Data|H_1)}$$

The subscript "1" in BF represents H and "0" represents H . Thus, BF denotes the Bayes factor for H versus H , while BF denotes H versus H . For example, BF = 10 means the probability of observing the current data under the null hypothesis is 10 times greater than under the alternative hypothesis. Based on Jeffreys' (1961) classification of different BF values, Wagenmakers et al. (2018) clarified the meaning of different BF magnitudes, though these guidelines are for reference only and researchers should evaluate BF meaning in the context of specific research questions.

Following Aczel et al. (2018), we used the BayesFactor R package developed by Morey et al. (2015) (function ttest.tstat) to calculate BF . The package' s default setting uses a two-sided Cauchy distribution as the prior for the alternative hypothesis (r = 0.707, where r is the scale parameter, also denoted as in some literature). Previous research suggests this prior setting for the alternative hypothesis is appropriate (Ly et al., 2016a, 2016b; Rouder et al., 2009). To examine the robustness of Bayes factor results, we also calculated BF using different prior distributions. One prior was a normal distribution (Dienes, 2014); compared to the default prior, the normal prior has relatively greater probability density near zero, yielding effects closer to zero than the default prior. Another prior was an informed prior based on expert opinion regarding effect size distributions (Gronau et al., 2019), reflecting experts' beliefs about effect size distributions (median = 0.350).

Given that researchers might mistakenly use p-values as evidence for the null hypothesis, we further explored the relationship between p-values and BF by calculating Kendall' s correlation coefficients (Kendall & Gibbons, 1990) and their corresponding 95% credible intervals (CIs) to assess whether p-values were strongly correlated with BF . If a strong correlation exists, larger p-values could to some extent support the null hypothesis; if no strong correlation exists, especially when p > 0.05 is weakly correlated with BF , using large p-values as evidence for the null hypothesis would be erroneous. Since the relationship is not linear, we used Kendall's . We employed the DescTools R package developed by Signorell (2017), using the function KendallTauB to calculate and the function credibleIntervalKendallTau (van Doorn et al., 2018) to compute 95% CIs based on and the number of t-tests. Finally, because larger sample sizes generally provide stronger evidence, we also explored the relationship between BF and sample size using the same method.

3. Results

3.1 Prevalence of Nonsignificant Results in Chinese Literature

Our analysis found that 36% of the 500 empirical articles contained at least one negative statement in their abstracts. Articles published in *Acta Psychologica Sinica* showed the highest proportion (43%), though all journals exceeded 30% (see Figure 2a [Figure 2: see original paper])[1]. These results indicate that negative statements are very common in psychological research.

Figure 2. (a) Proportion of negative statements across journals; (b) Proportion of negative statement interpretation categories across journals (Note: This classification is based on Interpretation ; see text regarding the two interpretations)

[1]Since the 500 articles included experimental, quasi-experimental, and survey research, the proportion of negative statements might differ across research types. We therefore analyzed the distribution of negative statements across research types within each journal. Results showed that experimental (45.8%) and quasi-experimental (36.2%) studies accounted for relatively larger proportions of negative statements than survey research (17.9%). However, because different journals emphasize different research directions, the proportion of research types varied considerably across journals. For example, in *Acta Psychologica Sinica*, *Psychological Science*, and *Studies of Psychology and Behavior*, experimental articles accounted for over 50% of publications. In contrast, *Chinese Journal of Clinical Psychology* and *Psychological Development and Education* contained larger proportions of survey and quasi-experimental research. Importantly, a single article could contain multiple studies, so we considered both cases where one study corresponded to multiple negative statements (e.g., Negative statement 1: "The results found no significant effect of Variable A on reaction time in Study 1" ; Negative statement 2: "The results found no significant effect of Variable B on reaction time in Study 1") and cases where one negative statement corresponded to multiple studies (e.g., "In both Study 1 and Study 2, no significant effect of Variable A on reaction time was found"). Therefore, the total number of negative statements considering research type was 301, exceeding the 236 negative statements mentioned earlier.

3.2 Classification of Negative Statements

Inter-rater agreement analysis for the six coders' classification of negative statements yielded a Fleiss' kappa of 0.588 ($p < 0.001$). Following Landis and Koch' s (1977) interpretation of Fleiss' kappa, this value indicates moderate inter-rater agreement. Additionally, authors discussed discrepant classifications to reach final results, making the classification reliable.

During classification, we frequently encountered statements like "no significant difference/effect/role" ($n = 55$). Due to ambiguity in Chinese expression, such statements allow two interpretations: Interpretation views them as direct in-

terpretations of $p < 0.05$, i.e., "the difference did not reach statistical significance," classified as "correct frequentist interpretation" ; Interpretation views them as descriptions supporting the null hypothesis, equivalent to "no difference/effect/role," classified as "incorrect frequentist interpretation—based on current sample." Therefore, in subsequent descriptions of classification results, we reported results based on both interpretations.

Classifying "no significant difference/effect/role"statements as correct frequentist interpretations, we categorized the 236 negative statements. Results showed that correct frequentist interpretations accounted for 53.4% (n = 126); incorrect frequentist interpretations accounted for 41.1% (n = 97), with 13.6% (n = 32) in the subcategory "incorrect frequentist interpretation—based on current sample" and 27.5% (n = 65) in "incorrect frequentist interpretation—generalization to population." Additionally, 5.5% (n = 13) were ambiguous and difficult to classify, coded as "difficult to judge." The distribution across categories for each journal is shown in Figure 2b. Classification results based on Interpretation [2] are in the footnote[2]. Aczel et al. (2018) also included a Bayes factor-based category, but we found no negative statements in this category—that is, no cases in these articles used Bayes factors to evaluate support for the null hypothesis—so we excluded this category from our study.

[2]If we classify "no significant difference/effect/role" statements as incorrect frequentist interpretation—based on current sample and reclassify the 236 negative statements, the change in interpretation only affects the counts for correct frequentist interpretation and incorrect frequentist interpretation—based on current sample, leaving other categories unchanged. Results show correct frequentist interpretation accounts for 30.1% (n = 71), while incorrect frequentist interpretation accounts for 64.4% (n = 152), with 36.9% (n = 87) in the subcategory incorrect frequentist interpretation—based on current sample and 27.5% (n = 65) in incorrect frequentist interpretation—generalization to population.

3.3 Bayes Factor Analysis

Within the NHST framework, researchers can only make binary decisions about rejecting the null hypothesis based on p-values and cannot obtain evidence supporting the null hypothesis. Therefore, we recalculated BF using t-test data to assess the degree to which data with nonsignificant results supported the null hypothesis.

Among all statistical tests, 39 reported t-test statistics and sample sizes. Based on these t-values and sample sizes, we calculated BF using a medium-scale two-sided Cauchy distribution as the prior for the alternative hypothesis. BF ranged from 0.51 to 10.64. Following Wagenmakers et al.' s (2018) classification of BF , we used 1, 3, and 10 as cutoffs to categorize BF into "weak evidence for H ," "weak evidence for H ," "moderate evidence for H ," and "strong evidence for H ." Results showed that among the 39 t-tests, 2.6% (n = 1) indicated weak evidence for H , 33.3% (n = 13) indicated weak evidence for H , 59% (n = 23)

indicated moderate evidence for H , and only 5.1% (n = 2) indicated strong evidence for H . In other words, if authors made inferences supporting H in the original articles, BF showed that only about half of these tests provided moderate or strong evidence for H . Therefore, inferring H as true based on p-values is inappropriate.

To verify result robustness and avoid prior specification effects, we recalculated Bayes factors using normal and informed priors. BF distributions under different prior settings are shown in Figure 3a [Figure 3: see original paper]. With the normal prior, BF ranged from 0.45 to 6.00; 15.4% (n = 6) indicated weak evidence for H , 64.1% (n = 25) indicated weak evidence for H , and 20.5% (n = 8) indicated moderate evidence for H . With the informed prior, BF ranged from 0.41 to 21.69; 20.5% (n = 8) indicated weak evidence for H , 53.8% (n = 21) indicated weak evidence for H , 17.9% (n = 7) indicated moderate evidence for H , and only 7.7% (n = 3) indicated strong evidence for H .

Figure 3. (a) Distribution and interpretation of BF under different prior settings; (b) Relationship between BF and p-values under default prior; (c) Relationship between BF and sample size under default prior

Note: Multiple BF values may correspond to the same sample; for example, a sample size of 138 corresponds to multiple BF values.

We further examined how prior settings affected negative statement classification. Results showed that changing from the default prior to the informed prior altered the interpretation of BF in 60% of cases (n = 23); changing to the normal prior altered interpretation in 61.5% of cases (n = 24). This indicates that prior distribution settings substantially influence BF interpretation, and researchers must carefully select appropriate priors when calculating BF .

Finally, we conducted exploratory analyses of Bayes factors, examining relationships between BF and p-values and between BF and sample size. Since only 39 t-tests reported t-statistics and sample sizes, our correlation results should be considered preliminary and require verification in future research. To explore the relationship between p-values and corresponding BF , we plotted p-values against BF (Figure 3b) and calculated Kendall' s and its 95% credible interval. Results showed a correlation of 0.527 between p-values and BF , with a 95% CI of [0.282, 0.687], indicating that larger p-values correspond to larger BF values. However, Figure 3b shows that this relationship is primarily driven by smaller p-values ($p < 0.2$); as p-values increase, BF changes become more gradual. Therefore, the validity of this conclusion requires further investigation.

Similarly, we analyzed the relationship between sample size and BF (Figure 3c). Results showed a correlation of 0.243 between sample size and BF , with a 95% credible interval of [0.018, 0.431], indicating a weak relationship. Figure 3c also shows that BF does not change substantially with increasing sample size. However, sample sizes ranged primarily from 13 to 138, with only a few studies exceeding 300; thus, the accuracy of this conclusion awaits further verification.

4. Discussion

This study analyzed 500 randomly selected Chinese psychology empirical research articles, extracting all negative statements from abstracts and recalculating Bayes factors using associated statistics and sample sizes to investigate the prevalence and accuracy of nonsignificant result interpretation in Chinese psychology core journals and compare this with international journals.

Regarding the prevalence of negative statements, we found that 36% of article abstracts (n = 180) contained negative statements in which researchers explicitly stated that an effect was absent or mentioned nonsignificant results. For example, over 40% of empirical articles published in *Acta Psychologica Sinica* contained negative statements in their abstracts. In contrast, Aczel et al. (2018) reviewed empirical studies published in international core journals (*Psychonomic Bulletin & Review*, *Journal of Experimental Psychology: General*, and *Psychological Science*) and found that 32% of abstracts mentioned negative statements—a proportion lower than our survey of Chinese journals. Combined with Aczel et al. (2018), our results demonstrate that nonsignificant results are indispensable in psychological research, as researchers need them to support their inferences, especially in experimental studies (the primary research type analyzed by Aczel et al., 2018), where negative statements appeared in 45.8% of cases.

Regarding accuracy of nonsignificant result interpretation, although 41.1% of statements contained misinterpretations (treating "no significant difference" as correct phrasing, i.e., Interpretation), Aczel et al. (2018) found that 72% of articles in international journals misinterpreted nonsignificant results. Even when we treated the common Chinese phrasing "no significant difference" as incorrect (under Interpretation ²: 64.4%), the misinterpretation rate remained lower than in international journals. These results indicate that while Chinese researchers, like their international peers, commonly misinterpret nonsignificant results, the proportion of misinterpretations appearing in articles is lower than in international psychology journals. Notably, classification results differed by over 20% depending on interpretation, suggesting researchers should use clear and explicit phrasing for statistical inference statements.

Furthermore, Bayes factor analyses revealed that few BF values exceeded 10 regardless of prior distribution (default prior: n = 2; normal prior: n = 0; informed prior: n = 3), with most BF values below 3 (default prior: n = 14; normal prior: n = 31; informed prior: n = 29). Although researchers may differ in their interpretation of BF evidence strength (Schönbrodt, 2015), most interpret BF < 3 as weak evidence for the null hypothesis and BF > 10 as strong evidence (Lee & Wagenmakers, 2014). These Bayes factor results are similar to those from international journals, though the small sample sizes providing stronger evidence prevent us from concluding that Chinese journals have a clear advantage on this point. Bayes factor analyses show that BF calculated from data with nonsignificant results rarely provides strong evidence for the null hypothesis. However, most t-tests in our Bayes factor analysis

had sample sizes under 100; Aczel et al. (2018) suggested this result may partly reflect small sample sizes common in psychological research (Button et al., 2013; Stussi et al., 2018; Cui et al., 2019; Xie et al., 2019). Hoekstra et al. (2018) reanalyzed nonsignificant results in medical research and found that with large sample sizes, data provided strong support for the null hypothesis.

We also explored correlations between p-values, sample size, and BF . However, our correlation analyses were preliminary, involving only 39 t-tests, so we encourage future research to examine these relationships more thoroughly. For p-values and BF , the correlation coefficient was 0.527. Similar to Aczel et al. (2018), we found that the positive correlation between p-values and BF occurred mainly among nonsignificant results with smaller p-values. When p < 0.2, BF increased with p-values; however, when p-values were larger, increases in p-values did not substantially affect BF . This reflects NHST' s limitation: p-values do not have clear meaning, cannot measure the probability that a research hypothesis is true or false, and larger p-values do not indicate stronger evidence for the null hypothesis (Hao et al., 2016; X. Lyu et al., 2020). Wetzels et al. (2011) similarly found that when p-values are large, BF changes little with p-values. Beyond psychology, Hoekstra et al. (2018) analyzing nonsignificant results in medical research found a linear relationship between log(BF) and p-values, with BF changes diminishing as p-values increased. For sample size and BF , the correlation was only 0.243, indicating that BF changes little with increasing sample size. Yet p-values are affected by sample size (Cheng & Li, 2019): even with very small effect sizes, sufficiently large samples can easily produce significant results. Therefore, research conclusions should not focus solely on whether statistical results are significant but should combine statistical results with practical significance.

However, as noted earlier, the number of t-tests and corresponding sample sizes in our Bayes factor analysis were small, so the generalizability of these results awaits verification.

Notably, Aczel et al. (2018) found that 10% of negative statements used Bayes factors rather than NHST for statistical inference. In contrast, none of the 500 articles we randomly selected used Bayes factors, reflecting that Chinese researchers are less familiar with methods that can support the null hypothesis. Lü (2012) therefore recommends that researchers pay more attention to alternative statistical inference methods as supplements to NHST, introducing the principles behind different methods at an appropriate level of difficulty to enable more comprehensive understanding of their advantages and disadvantages —such as equivalence testing (Lakens et al., 2018; Lakens et al., 2018; Rogers et al., 1993), Bayesian estimation (Kruschke, 2011; Kruschke & Liddell, 2018; McElreath, 2018), and Bayes factors (Wagenmakers et al., 2018; Wagenmakers et al., 2011; Hu et al., 2018). For specific methodological guidance, see Lu et al. (2020).

An important difference between this study and Aczel et al. (2018) is our additional "difficult to judge" category. For example, article 2052 stated: "In the

generalization task, pain expressions only prolonged subjective duration under the second-level condition,"which implicitly suggests that no effect existed under other conditions or that no effect was found under other conditions—corresponding to incorrect and correct interpretations, respectively. However, we could not determine the authors' intended meaning, so we classified such descriptions as "difficult to judge." Such ambiguous phrasing partly reflects researchers' neglect of statements about nonsignificant results, focusing only on significant findings. We also found nonstandard terminology, such as "In athlete populations, high state anxiety had little effect on processing efficiency and accuracy," further indicating that researchers need to treat nonsignificant results more carefully.

Although this study reveals misinterpretation of nonsignificant results in current literature, it cannot explore the causes of these misunderstandings. One possible cause may be erroneous p-value interpretations in textbooks. For example, Cassidy et al. (2019) surveyed North American psychology textbooks and found that many contained misinterpretations of p-values. Chinese textbooks also contain incorrect interpretations of nonsignificant results. For instance, Zhang and Xu (2015) wrote in Chapter 8: "The problem of hypothesis testing is to determine whether the null hypothesis H is correct, deciding to accept or reject the null hypothesis H " ; Lu (2009) wrote in Chapter 7: "If, under the condition that the null hypothesis H holds, the probability of a statistic calculated from the sample is not very small, then accept the null hypothesis." Both statements suggest that NHST can provide evidence for accepting the null hypothesis, which may contribute to Chinese researchers' misinterpretation of nonsignificant results.

This study has several limitations. First, with 13 coders, there may have been differences in understanding the coding manual, such as inconsistent lengths of extracted negative statements, suggesting coders may have differed in their understanding of classification criteria. To minimize these differences, each article was coded by at least two coders, with the second verifying the first' s work. For classification of negative statements—the study' s focus—six coders first worked independently, then discussed discrepancies, achieving moderate inter-rater agreement as assessed by Fleiss' kappa, indicating reliable coding results. Second, when quantifying evidence for the null hypothesis using Bayes factors, we only used t-test data, excluding many other statistical tests such as correlation coefficients. However, our results are consistent with Aczel et al. (2017), who recalculated Bayes factors for significant results involving t-tests, F-tests, and correlation analyses across 35,515 published articles and found similar evidence strength across different statistical tests in psychological research, suggesting our t-test-based results may generalize to other tests. Third, we only included data from 2017 and 2018, reflecting the situation only during that period and providing no information on trends over the past five or ten years. Fourth, in clinical trials, erroneously accepting the null hypothesis may lead to incorrect inferences about group matching on certain variables, but such detailed information generally does not appear in abstracts; future research could conduct full-text searches on this issue.

Despite these limitations, our findings suggest that researchers in psychology and other empirical sciences need to re-examine conclusions corresponding to nonsignificant results. Misinterpreting nonsignificant results can have serious consequences: overlooking actual differences between experimental and control groups in between-subjects designs and ignoring true effects that may be masked by nonsignificant results in small-sample studies (Jia et al., 2018). Misinterpretation of nonsignificant results may also contribute to publication bias (Franco et al., 2014; Kühberger et al., 2014), potentially inducing p-hacking behavior (Head et al., 2015) and leading to non-replicable research or severely reduced effect sizes (Baker, 2016; Klein et al., 2014; Open Science Collaboration, 2015; Hu et al., 2016). Researchers should therefore strengthen the rigor of their nonsignificant result interpretations to avoid these negative consequences.

5. Summary

By analyzing 500 empirical studies from five Chinese psychology journals, this study found that negative statements are relatively common in Chinese literature, with a higher proportion than in international journals, indicating that nonsignificant results hold an important position in psychological empirical research. Regarding interpretation of nonsignificant results, the misinterpretation rate in Chinese journals was lower than in international journals. Additionally, Bayes factor analysis showed that data with nonsignificant results in the literature could not provide strong evidence supporting the null hypothesis. Overall, Chinese researchers need to strengthen their understanding of nonsignificant results and use appropriate statistical methods to evaluate the degree of support for the null hypothesis, thereby reducing misinterpretation of nonsignificant results and improving the quality of psychological research.

References

Cheng, K. M., & Li, S. E. (2019). P-values in scientific research: Misunderstandings, manipulation, and improvements. *Journal of Quantitative and Technical Economics*, *7*, 117–136.

Cui, Y. C., Wang, P., & Cui, Y. J. (2019). Cognitive control strategies in perceptual conflict impression formation: Using stereotypical and counter-stereotypical information as examples. *Acta Psychologica Sinica*, *51*(10), 1157–1170.

Hao, L., Liu, L. P., & Shen, Y. F. (2016). Statistical significance: A misunderstood p-value. *Statistics and Information Forum*, *31*(12), 3–10.

Hu, C. P., Kong, X. Z., Wagenmakers, E.-J., Ly, A., & Peng, K. P. (2018). Bayes factors and their implementation in JASP. *Advances in Psychological Science*, *26*(6), 951–965. doi:10.3724/SP.J.1042.2018.00951

Hu, C. P., Wang, F., Song, M. D., Sui, J., & Peng, K. P. (2016). The reproducibility crisis in psychological research: From crisis to opportunity. *Advances in Psychological Science*, *24*(9), 1504–1518. doi:10.3724/SP.J.1042.2016.01504

Lu, S. H. (2009). *Social statistics* (4th ed.). Peking University Press.

Lu, C. L., Wang, J., Song, Q. Y., Jia, B. B., Xu, Y. P., & Hu, C. P. (2020). Methods for extracting information from nonsignificant results: Principles and implementation. [ChinaXiv:202001.00113].

Luo, D. S. (2017). Evaluating two roots of psychology' s reproducibility crisis. *Studies of Psychology and Behavior*, *15*(5), 577–586.

Lü, X. K. (2012). The Fisher-Neyman-Pearson controversy and debates over hypothesis testing in psychological statistics. *Psychological Science*, *35*(6), 1502–1506.

Lü, X. K. (2014). From tool to paradigm: A sociological reflection on the hypothesis testing controversy. *Chinese Journal of Sociology*, *34*(6), 216–236.

Sun, H. W., Dong, Z. J., & Zhao, Y. J. (2012). Misunderstandings and misuses of statistical hypothesis testing. *Chinese Journal of Health Statistics*, *29*(1), 147–148.

Xie, S. S., Zhang, J. J., & Zhu, J. (2019). Categorical perception of color occurs in both cerebral hemispheres: Evidence from Naxi and Han participants. *Acta Psychologica Sinica*, *51*(11), 1229–1243.

Zhang, H. C., & Xu, J. P. (2015). *Modern psychological and educational statistics* (4th ed.). Beijing Normal University Press.

Zhong, X. B. (2016). The hypothesis testing controversy: Clarification and resolution. *Advances in Psychological Science*, *24*(10), 1670–1676.

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

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ···Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. doi:10.1177/251524591877374

Algermissen, J., & Mehler, D. M. (2018). May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology*, *119*(6), 2114–2117. doi:10.1152/jn.00765.2017

American Psychological Association. (2010). *Publication manual of the American Psychological Association*. Washington, DC: American Psychological Association.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. doi:10.1038/d41586-019-00857-9

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. doi:10.1038/533452a

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233–239. doi:10.1177/2515245919858072

Chen, X., Lu, B., & Yan, C.-G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping*, *39*(1), 300–318. doi:10.1101/128645

Chuard, P. J., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLOS Biology*, *17*(1), e3000127. doi:10.1371/journal.pbio.3000127

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 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*(3), 193–242. doi:10.1037/h0044139

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. doi:10.1007/s11192-011-0494-7

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from -error control to validity proper. *Perspectives on Psychological Science*, *7*(6), 661–669. doi:10.1177/1745691612462587

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. doi:10.1126/science.1255484

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (R package version 0.84.1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=irr

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 392–410). Thousand Oaks, CA: SAGE Publications.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi:10.1007/s10654-016-0149-3

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, *73*(S1), 262–269. doi:10.1080/00031305.2018.1562983

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106. doi:10.1371/journal.pbio.1002106

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), e0195474. doi:10.1371/journal.pone.0195474

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Jia, X.-Z., Zhao, N., Barton, B., Burciu, R., Carriere, N., Cerasa, A., ···Zang, Y.-F. (2018). Small effect size leads to reproducibility failure in resting-state fMRI studies. *bioRxiv.* doi:10.1101/285171

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ···Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. doi:10.1177/1745691611406925

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-4

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, *9*(9), e105825. doi:10.1371/journal.pone.0105825

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. doi:10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian reanalyses from summary statistics: A guide for academic consumers. *Advances in Methods and Practices in Psychological Science*, *1*(3), 367–374. doi:10.1177/2515245918779348

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. doi:10.1016/j.jmp.2016.01.003

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys' s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi:10.1016/j.jmp.2015.06.004

Lyu, X., Xu, Y., Zhao, X., Zuo, X.-N., & Hu, C.-P. (2020). Beyond psychology: The prevalence of misinterpretation of p-values and confidence intervals across different fields. *Journal of Pacific Rim Psychology*, *14*, e6. doi:10.1017/prp.2020.6

Lyu, Z., Peng, K., & Hu, C.-P. (2018). P-value, confidence intervals and statistical inference: A new dataset of misinterpretation. *Frontiers in Psychology*, *9*, 868. doi:10.3389/fpsyg.2018.00868

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Miller, G. (2011). ESP paper rekindles discussion about statistics. *Science*, *331*(6015), 272–273. doi:10.1126/science.331.6015.272

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

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi:10.1037/1082-989X.5.2.241

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:10.1037/0033-2909.113.3.553

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*(2), 225–237. doi:10.3758/PBR.16.2.225

Schäfer, T., & Schwarz, M. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. doi:10.3389/fpsyg.2019.00813

Schönbrodt, F. (2015). 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

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General*, *147*(6), 905–917. doi:10.1037/xge0000403

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall' s rank correlation coefficient. *The American Statistician*, *72*(4), 303–308. doi:10.1080/00031305.2016.1264998

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ···Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. doi:10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi:10.1037/a0022790

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA' s statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t-tests. *Perspectives on Psychological Science*, *6*(3), 291–298. doi:10.1177/1745691611406923

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance.* Ann Arbor: University of Michigan Press.

Note: Figure translations are in progress. See original paper for figures.