Review

What information is provided from non-significant findings and how can this be improved?

Scott J. Dankel

Objectives: To clarify what information is provided from non-significant findings and explain possible additional/alternative tests to help make these findings more informative.

Design & Methods: The design of this manuscript was to first clarify what information is provided from non-significant findings and detail why this may be different than what is commonly thought. Next, information is given as to why it may be particularly important for non-significant findings to be further examined within the field of exercise science given that small sample sizes are often employed. Lastly, a brief overview of two possible ways in which researchers can make non-significant findings more informative is provided.

Results & Conclusions: Non-significant findings alone do not provide strong support that a given intervention did not have an effect. Researchers may wish to instead use a Bayesian statistical approach capable of quantifying evidence for both the null and alternative hypotheses. For researchers who prefer to use frequentist statistical approaches, a test for statistical equivalence may be used when there is no statistical difference present. These approaches may provide more insight into whether non-significant findings are due to uncertainty in the data or support for the null hypothesis.

(*Journal of Trainology* 2019;8:19-23)

Key words: Bayesian ■ null findings ■ publication bias ■ replication crisis ■ statistical equivalen

INTRODUCTION

A publication bias exists in the scientific literature with approximately 86% of studies published in 2007 demonstrating statistically significant results.[1] Interestingly, the acceptance rate among articles submitted to medical journals does not differ drastically based on whether non-significant (15.0%) or statistically significant findings (20.4%) are reported[2]. Therefore, the scarcity of published studies containing non-significant findings appears to be largely related to a reluctance of authors to submit such manuscripts for publication. This has led to a push for authors to submit non-significant findings,[3,4] with some journals (e.g. Journal of Articles in Support of the Null Hypothesis, Journal of Negative Results in BioMedicine and Journal of Pharmaceutical Negative Results) directly catered toward publishing such studies[5]. Given the push toward publishing non-significant findings, it is important to understand what information these studies actually provide.

What information do non-significant findings provide?

In the scientific literature, the majority of studies are analysed using frequentist statistical approaches where probabilities (i.e. p-values) are obtained. These statistical approaches provide information on the probability of the observed test statistic assuming the null hypothesis is true.[6] In other words, any p-values less than 0.05 indicates that the observed effect was large enough (i.e. a large mean difference) and/or consistently observed (i.e. low variability) to such an extent that the effect would be unlikely the result of random chance (i.e. less than 5%). It is important to note that the statistical tests employed answer the following question: Is the magnitude of the effect great enough to where it would be unlikely to be the result of random chance? While p-values are often used as support for the alterative hypothesis, p-values do not provide information on the probability that the alternative hypothesis is *true*.[7]

So what happens when the p-value is greater than 0.05? This is often misinterpreted as evidence that there is no effect, but failing to reject the null hypothesis does not provide evidence in favor of the null hypothesis.[8] So while the p-value can provide some evidence for the alternative hypothesis, the p-value alone cannot provide evidence for the null hypothesis. There are usually two possible conclusions that researchers make from a research study, and these are largely impacted by the p-value. That is, researchers will either conclude there is an effect of the intervention when the p-value is less than 0.05, or that there is not an effect of the intervention when the p-value is greater than 0.05. Unfortunately, researchers often dichotomize the p-value and only conclude that there either is or is not an effect and leave out the possibility of uncertainty.[9] A third possible conclusion that can be made when the p-value is greater than 0.05 is that there is not enough information to make a conclusion based on the data. When only p-values are reported it is not possible to quantify evidence for the null hypothesis, and therefore, studies reporting non-significant findings do not necessarily indicate the absence of an effect.[8] It is entirely plausible that there is too much error surrounding the response (i.e. a wide confidence

Received June 29, 2019; accepted August 1, 2019

From the Department of Health and Exercise Science, Exercise Physiology Laboratory, Rowan University, Glassboro, NJ, USA

Communicated by Takashi Abe, PhD

Correspondence to: Dr. Scott J. Dankel, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ 08028, USA. Email: dankel47@rowan.edu

interval) that a definitive conclusion cannot be made. Consider a weight loss intervention that produces a 0 kg change in body mass. If the intervention produced a mean weight loss of 0 kg with a 95% confidence interval between -0.2 kg and 0.2 kg the reader can be fairly confident that the intervention had little effect. On the contrary, if the mean weight loss is 0 kg with a 95% confidence interval between -15 kg and 15 kg it would be hard to confidently conclude that the intervention had no effect since the precision of the estimate is poor and includes the possibility of large clinically meaningful losses in body mass. Notably, both of these findings would result in a non-significant p-value for a paired t-test. Therefore, non-significant findings (i.e. non-significant p-values) alone should not be used as support that a given intervention had no effect.

Underpowered studies limit the interpretability of non-significant findings

One of the reasons why it may be particularly important to further assess non-significant findings is because many studies involving human participants have inadequate power (i.e. the sample sizes are too small).[10,11] In statistics there are two main types of errors that can be made. The alpha level is the probability of a type 1 error and occurs when the researcher concludes there is an effect when there is not one (i.e. a false positive). The beta level is the probability of a type II error and occurs when the researcher concludes there is not an effect when there is one (i.e. a false negative). The power of a study is the probability that an effect will be detected if there is one. Therefore, statistical power will be inversely proportional to the type II error rate since the type II error rate is the probability that an effect was not detected. Thus, power is calculated as 1 minus the type II error rate. When researchers conduct a study they obtain a set number of individuals and end up with a given effect size from the study. Since researchers almost always use an alpha level (type I error rate) of 0.05, it is the beta level (type II error rate) that is free to fluctuate and increases as a result of small sample sizes. For example, some estimations of the average study power across different fields demonstrated an estimated power of 0.35 (65% type II error rate) in psychological research and 0.21 (79% type II error rate) in neuroscience research.[11] This is an extremely high type II error rate especially considering some researchers have suggested setting both the alpha and beta levels to 0.05.[12] There are consequences of both type I and type II errors. For example, if a new supplement is being tested, a type I error may result in consumers purchasing a product that they think works when in fact it does not. Contrarily, a type II error may result in dismissing a supplement as ineffective when it could in fact be beneficial for consumers.

To put into perspective how underpowered some studies may be, a study comparing pre to post changes across two groups (intervention and control), with an estimated moderate effect (d = 0.5), would require 128 individuals (64 per group) if an alpha level of 0.05 and power of 0.8 are used. Using the same effect size and alpha level, a more commonly employed sample size of 20 individuals (10 per group) would yield a power of 0.18, therefore drastically inflating the type II error rate from 20% to 82%. Thus, non-significant findings may be the result of a truly ineffective intervention or they may be the result of inadequate power. Put simply, many studies within the exercise science literature are underpowered which limits the interpretability of non-significant findings.

Making non-significant findings more interpretable

There are ways by which researchers can determine whether non-significant findings are the result of uncertainty or whether they truly provide support for the null hypothesis. Aside from increasing the sample size to help reduce type II error rates, there are additional or alternative statistical analyses that can be done once the study is complete. Two possible approaches include either using a Bayesian analysis or computing an equivalence test to follow-up non-significant findings. While this manuscript is not intended to provide a comprehensive guide on how to conduct these analyses, a brief overview is provided. Bayesian statistical analyses can be performed using the JASP software[13] and tests of statistical equivalence can be performed using the TOSTER equivalence testing package in Jamovi[14]. Both of these software packages are point and click and very user friendly. These additional/alternative analyses may help to provide more insight into what information non-significant findings may actually be providing (Figure 1).

Using a Bayesian Approach

A Bayesian analysis is unique to that of traditional frequentist statistical approaches in that evidence can be provided toward the null hypothesis.[15] Bayesian analyses require inputting prior odds which consist of an estimated effect size and variance before the data is even collected. The prior odds are then multiplied by a Bayes Factor in order to obtain the posterior odds, which provides an effect size and variance measure after taking into account the collected data. The focus of this manuscript will be on the Bayes Factor as this can be used to provide evidence for the null or alternative hypothesis. Bayes Factors are interpreted just like odds ratios, such that a Bayes Factor of 10, for example, would indicate that the alternative hypothesis is 10 times more likely than the null hypothesis. On the other hand a Bayes Factor of 0.1, for example, would indicate that the null hypothesis is 10 times more likely than the alternative hypothesis. Support for the null hypothesis can be interpreted as there not being any effect of the intervention. On the contrary, an odds ratio of 1, for example, would indicate that the null hypothesis is equally as likely as the alternative hypothesis, and thus this provides little information on the efficacy (or lack thereof) of the intervention. In other words, although there is not support for the alternative hypothesis there is also not support for the null hypothesis and more data is necessary to make a strong conclusion on the efficacy of the intervention. This information cannot be obtained from a p-value provided from a traditional frequentist statistical test. One benefit of the Bayesian approach is that, if a non-informative Bayes Factor is present after running the analysis (i.e. the Bayes Factor is close to 1 indicating

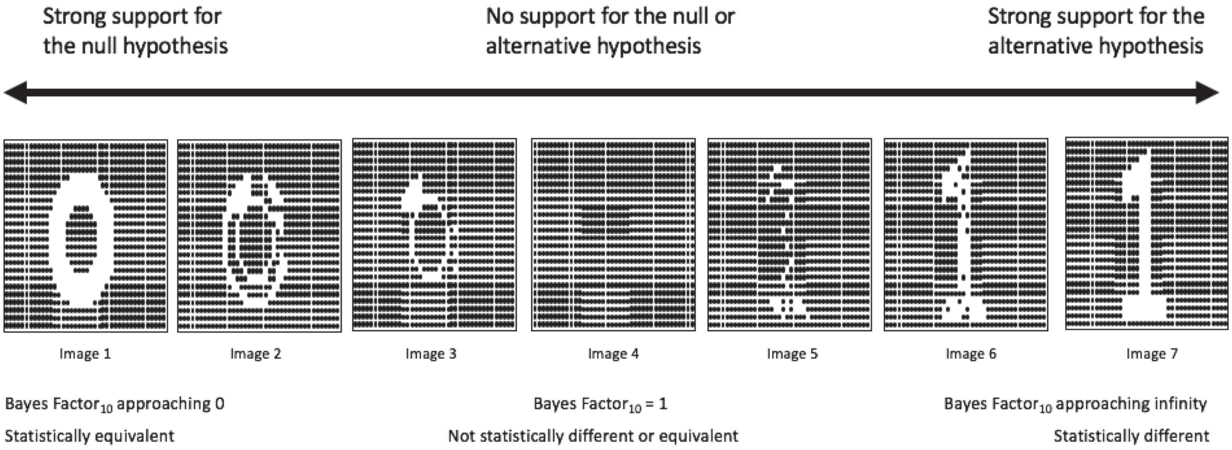
Strong support for the null hypothesis
No support for the null or alternative hypothesis
Strong support for the alternative hypothesis
Image 1
Image 2
Image 3
Image 4
Image 5
Image 6
Image 7
Bayes Factor$_{10}$ approaching 0
Statistically equivalent
Bayes Factor$_{10}$ = 1
Not statistically different or equivalent
Bayes Factor$_{10}$ approaching infinity
Statistically different

Figure 1 An illustration of possible statistical tests that may give more insight into the information provided by non-significant findings. A given research study may provide some insight into whether the data provides support for the null or alternative hypothesis. The "0" indicates support for the null hypothesis and the "1" indicates support for the alternative hypothesis. The clarity of the number indicates how much support the study provides. It is possible for a study to provide little information to the field (image 4) if the Bayes Factor indicates that the null hypothesis fits the data equally as well as the alternative hypothesis (a Bayes Factor of 1); or, using frequentist statistics, there is not a statistical difference or statistical equivalence. Simply running a traditional frequentist statistical test of differences will not allow the reader to see whether the data truly provides support for the null hypothesis (images 1-3) or whether the data is ambiguous (image 4). This figure illustrates that not all non-significant findings provide the same information. The Bayes Factor is reported with the subscript "10" to indicate that support for the alternative hypothesis is the numerator and support for the null hypothesis is the denominator. It has been suggested that less than 3 times greater support for one hypothesis over the other represents weak evidence, 3-10 times greater support for one hypothesis represents moderate evidence, and greater than 10 times support for one hypothesis represents strong evidence.[17]

there is similar support for both the null and alternative hypotheses), more data can be collected without impacting the interpretation of the Bayes Factor (whereas this would increase the alpha rate in frequentist statistical approaches).[16] One limitation with Bayesian statistics is that the results can be impacted by the prior that is chosen before data is collected. For this reason, researchers may wish to conduct a Bayes Factor robustness check to see how stable the Bayes Factor is across a wide range of prior widths.[17] Another limitation exists in that evidence is quantified along a continuum which may result in there being some subjectivity when interpreting results. This, however, can also be viewed as a positive in being able to quantify the strength of evidence present.

Using an equivalency test

Another possible approach to follow up non-significant findings is to use an equivalence test. This will test if the intervention is statistically equivalent to that of a control group (or equivalent to zero if a control group is not used). The idea behind equivalency testing is very similar to that of testing for statistical differences. For example, if there is only an intervention group, a traditional paired t test (using a common 0.05 alpha level) would test if the 95% confidence interval surrounding the mean does not cross zero. On the contrary a test of statistical equivalence would test if the 90% confidence interval surrounding the mean (90% because there is an upper and lower bound as opposed to just zero, and a 90% confidence interval yields a 0.05 alpha level for equivalence testing[18]) lies within the established boundaries. The established boundaries can be chosen as a value that would be clinically meaningful; or, a common rule of thumb is to use half of the minimal difference above and below zero[19] (i.e. the minimal difference centered on zero). The basic idea behind equivalency testing is to see if the 90% confidence interval around the observed effect contains any values that would be deemed clinically meaningful. If it does not, this indicates that the mean response is not meaningful, and there is not a large degree of variability (i.e. error) surrounding this response either. Collectively, if the 95% confidence interval were to lie outside of zero (i.e. not include zero) it would be statistically significant, if the 90% confidence interval were to lie within the established boundaries (i.e. not cross either boundary) it would be statistically equivalent, if neither of the two were present this would represent ambiguity in that there is not statistical equivalence nor is there a statistical difference. One possible limitation with using statistical equivalence testing is that a researcher could end up with a result that is both statistically different and statistically equivalent,[20] particularly if the established boundaries for equivalence are too wide or the study is overpowered. This could not happen with Bayesian statistics since the Bayes Factor is simply a ratio of evidence for the null and for the alternative hypotheses. Another limitation exists in that the results of an equivalence test will be largely influenced by the sample size, such that larger sample sizes will increase the likelihood of finding statistical equivalence by reducing the standard error and shrinking the 90% confidence interval. This same limitation exists for testing statistical differences.

An example using previous data

My previous laboratory members and I published a study in which we tested to see if pooling metabolites post-exercise would augment adaptations to resistance exercise.[21] We reported that there was no difference in strength gains when comparing traditional exercise (i.e. a control condition) to an experimental condition in which metabolites were pooled post-exercise. As we did not explore this non-significant finding further using either a Bayesian analysis or equivalence test, it may be difficult for the reader to determine if the lack of significance was truly due to an equivalent change in strength across both exercise conditions. Conducting a Bayesian analysis using JASP (JASP Team, 2019) with an uninformed prior of 0.707 cantered on zero, yields a Bayes Factor of 0.330 showing moderate support for the null hypothesis. Conducting a test of statistical equivalence using Jamovi (The Jamovi Project, 2019) with equivalence bounds of -1.45 and 1.45 kg (half of our time matched minimal difference of 2.9 kg) would demonstrate that the strength gains were not statistically equivalent (lower bound: $p = 0.07$, upper bound: $p = 0.008$). This is because the 90% confidence interval on the difference between measures was -1.62 to 0.77 kg, and the lower limit of -1.62 kg exceeds the boundary of -1.45 kg. Thus, while close, the strength gains were not statistically equivalent. This example may help to show the subjective nature of equivalence testing, because had we chosen boundaries of 1.65 kg instead of 1.45 kg, the results would be statistically equivalent.

What to do if there is no support for the null or alternative hypothesis

For researchers, it is best if the data provides support for either the null or the alternative hypothesis to indicate that the intervention either did or did not have an effect. If neither the null or alternative hypothesis is supported, the researcher is left to conclude that the study provides little to no information on the efficacy of the intervention. Unfortunately, this may deter some researchers from using a Bayesian or statistical equivalence approach as researchers may need to increase their sample sizes to increase the precision of their estimates in order to make a conclusion on the efficacy of a given intervention. Unless an equivalence test or Bayesian analysis is performed to provide support for the null hypothesis, researchers should not interpret non-significant findings as the absence of an effect, particularly when smaller sample sizes are used. Small sample sizes may be at the heart of the reproducibility crisis,[22] because of the lack of power and inflated type II error rate discussed previously. Therefore, one study may find a significant effect and another study may not, but this does not mean that the study not finding an effect was the results of support for the null hypothesis (e.g. these two studies may be depicted as images 5 and 6 in Figure 1). Again, the absence of statistical significance may be the result of insufficient power in which case there is not support for the alternative hypothesis nor is there support for the null hypothesis. If one study provides an ambiguous result, this should not be considered a failure to replicate a previous finding, because an ambiguous result does not provide any information for or against the efficacy of the intervention. A truly non-replicable study would be the result of one study providing support for the alternative hypothesis and another providing support for the null hypothesis (e.g. images 1 and 7 in Figure 1). This again highlights the importance testing both the null and alternative hypotheses, as opposed to just testing the alternative hypothesis.

CONCLUSION

Non-significant findings alone do not provide strong support that a given intervention did not have an effect, particularly given that many studies use small sample sizes that are underpowered. Unless there is support for the null hypothesis that there truly is no effect of the intervention (i.e. it is equivalent to zero or equivalent to a control group response), non-significant findings should not automatically be interpreted as support for the null hypothesis. Future studies may wish to employ Bayesian statistical approaches which can quantify evidence for or against both the null and alternative hypotheses. Researchers who prefer frequentist statistical approaches may wish to tests for statistical equivalence when there is an absence of statistical significance (i.e. there is no statistical difference).

CONFLICT OF INTEREST

None

REFERENCES

1. Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics* 2012;90:891-904.
2. Olson CM, Rennie D, Cook D et al. Publication bias in editorial decision making. *JAMA* 2002;287:2825-2828.
3. Sandercock P. Negative results: why do they need to be published? *Int J Stroke* 2012;7:32-33.
4. Weintraub PG. The Importance of Publishing Negative Results. *J Insect Sci* 2016;16(1) available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088693/. doi:10.1093/jisesa/iew092.
5. Mlinarić A, Horvat M, Šupak Smolčić V. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochem Medica* 2017;27:30201.
6. Verdam MGE, Oort FJ, Sprangers MAG. Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. *Qual Life Res* 2014;23:5-7.
7. de Graaf TA, Sack AT. When and How to Interpret Null Results in NIBS: A Taxonomy Based on Prior Expectations and Experimental Design. *Front Neurosci* 2018;12:915.
8. Harms C, Lakens D. Making "null effects" informative: statistical techniques and inferential frameworks. *J Clin Transl Res* 2018;3(Suppl 2):382-393.
9. de Graaf TA, Sack AT. Null results in TMS: from absence of evidence to evidence of absence. *Neurosci Biobehav Rev*. 2011;35:871-877.
10. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med*

1978;299:690-694.

11. Schweizer G, Furley P. Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychol Sport Exerc* 2016;23:114-122.
12. Oberhofer AL, Lennon RP. A call for greater power in an era of publishing negative results. *Acta Medica Acad* 2014;43:172-173.
13. JASP Team (2019). JASP (Version 0.10.2)[Computer software]. Retrieved from https://jasp-stats.org/
14. The jamovi project (2019). jamovi (Version 0.9) [Computer Software]. Retrieved from https://www.jamovi.org
15. Wagenmakers E-J, Marsman M, Jamil T et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev* 2018;25:35-57.
16. Rouder JN. Optional stopping: no problem for Bayesians. *Psychon Bull Rev* 2014;21:301-308.
17. van Doorn J, van den Bergh D, Bohm U et al. The JASP guidelines for conducting and reporting a Bayesian analysis. 2019. Available from: https://psyarxiv.com/yqxfr/
18. Walker E, Nowacki AS. Understanding Equivalence and Noninferiority Testing. *J Gen Intern Med* 2011;26:192-196.
19. Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis* 2008;66:150-154.
20. Lakens D. Equivalence Tests. *Soc Psychol Personal Sci* 2017;8:355-362.
21. Dankel SJ, Buckner SL, Jessee MB et al. Post-exercise blood flow restriction attenuates muscle hypertrophy. *Eur J Appl Physiol* 2016;116: 1955-1963.
22. Fraley RC, Vazire S. The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS One* 2014;9:e109019.