International Journal of EXERCISE SCIENCE

Editorial

# New Author Guidelines in Statistical Reporting: Embracing an Era Beyond p < .05

SAMANTHA L. JOHNSON‡1, WHITLEY J. STONE‡2, JENNIFER A BUNN‡3, T. SCOTT LYONS‡2, and JAMES W. NAVALTA‡4

[1]Department of Health and Human Performance, Middle Tennessee State University, Murfreesboro, TN, USA; [2]School of Kinesiology, Recreation, & Sport, Western Kentucky University, Bowling Green, KY, USA; [3]Department of Physical Therapy, Campbell University, Buies Creek, NC, USA; [4]Department of Kinesiology and Nutrition Sciences, University of Nevada, Las Vegas, Las Vegas, NV, USA

‡Denotes professional author

ABSTRACT

***International Journal of Exercise Science 13(1): 1-5, 2020.*** Statistical reporting of quantitative research data has been plagued by potential bias and reporting suppression due to a single numerical output: the *p*-value. While there is great importance in its merit, creating a pass-fail system (set at point of .05) has created a culture of researchers submitting their project's data to a filing cabinet if it does not yield "statistical significance" based on this value. The editors of the International Journal of Exercise Science are following the American Statistical Association's call for statistical reform by adjusting our reporting guidelines to the following requirements: [1.] make an intentional effort to move away from statements "statistically significant" or "not significant;" [2] all *p*-values are to be reported in their raw, continuous form; [3.] measures of the magnitude of effect must be presented with all *p*-values; [4.] either an a-priori power analysis with relevant citations should be included or post-hoc power calculations should accompany *p*-values and measures of effect. The ultimate goal of this editorial is to join with other scholars to push the field toward transparency in reporting and critical, thoughtful evaluation of research.

KEY WORDS: Statistical reform, alpha values, clinical significance, effect size, power analysis

## INTRODUCTION

The concept of statistical significance based on standard deviation was first addressed in 1885, with the primary intent of identifying data that should be analyzed further (6). However, by 1919 it became apparent there was a common misconception that statistical significance was a direct indicator of scientific importance (3). Although the concept was initially presented by Pearson in 1900 (12), the phrase "statistically significant" became customary following inclusion in a textbook by Fisher, which is also the publication that formalized and promoted the use of *p*-values (7). As the use of *p*-values and the phrase "statistically significant" became commonplace in reporting statistics, this concept was misinterpreted as a "go or no go" decisionmaking tool for researchers to determine importance of findings, bypassing the intended and necessary further analyses.

While the limitations of this statistical practice have been well known for at least 100 years, guidelines and educational practices for reporting statistical findings do not fully reflect this knowledge. Because of this, the American Statistical Association (ASA) has recently emphasized a move away from this statistical reporting practice (15). The purpose of this editorial is to address the implications of current reporting practice and identify ways the International Journal of Exercise Science (IJES) will reform policies and guidelines to align with recommendations provided by the ASA.

It is conventional for researchers in training to learn from quantitative psychologists on theory and methods for analyzing their investigations' data. While there is much to consider on the forefront of the study (how many participants are needed, the design of an intervention, and method to evaluate the data), too often minimal time is allocated to teach data output and *p* value interpretation in doctoral curriculum. Generally, the dogma of statistical significance is then applied, interpreting *p*-values less than .05 as "significantly different" from the other condition (control, baseline, other intervention, etc.), while those exceeding .05 are an indication that the values you are comparing are not "significantly different."

The concept of "statistical significance" being associated with a set threshold of .05 is described as an arbitrarily adopted practice stemming from Fisher's 1925 textbook (4-5, 7). This is not to belittle the work of Fisher, but merely to address the limitations of current practice. The presentation of using a binary cut-off for statistical significance at .05 was founded in a time where statistical findings were not a few clicks away on a computer, but rather, were calculated manually. Considering this, the value of .05 was practical and convenient, as this threshold suggests that the difference in values meets or exceeds approximately two standard deviations (7). Considering the resources available at the time, it is understandable how the practice of using .05 for determining significance became customary. However, Fisher (8) addressed that .05 may not be an appropriate criterion for all experiments, which is apparent based on the existence of more than one column in his tables. Furthermore, the use of the proposed threshold to identify significance is not an indication that it is the only needed analysis to establish scientific relevance. This is particularly important considering the vast expansion in resources from 1925 to present day. Today's scientists have a plethora of technological advantages in comparison to the researchers who first implemented this concept. Our methods of statistical reporting should reflect the progress in convenience and availability of calculating statistical values with great levels of acuity.

Considering the frequency of improper use and interpretation of *p*-values, the ASA has urged the scientific community to move away from the habit of simply reporting findings as "significant" or "non-significant" based on the *p*-value and a pre-determined alpha level (14). This binary mindset leads to selective reporting and publishing of manuscripts, which is detrimental to the quality of literature and increases the likelihood of bias in knowledge gained from published research (13, 15). It is not uncommon for researchers to file away entire studies because of "non-significant" findings as publishers are less inclined to print research that failed to identify a difference compared to a flashy, more "impactful" study. As such, IJES is releasing this Editorial as a guideline for our new statistical reporting expectations.

With the goal of improving the quality of literature available for scholars, IJES will begin implementing best practice recommendations from the ASA in statistical reporting. These changes will include [1.] ceasing the use of statements related to "statistical significance;" [2.] reporting of continuous *p*-values for all tested hypotheses; [3.] providing effect size for all tested hypotheses; and [4.] including either an a-priori or post-hoc power analysis with relevant citations.

## PUBLICATION EXPECTATIONS

### Eliminating Binary Statistical Decisions

Regardless of the intentions of the researcher, reporting data as simply "significant" or "non-significant" using a pre-determined threshold (commonly .05 in our field) can lead researchers and policy-makers to make decisions based solely on *p*-values rather than considering the full scientific importance of the findings (14). Furthermore, this practice can result in pressure for journals and reviewers to prioritize publishing manuscripts with findings classified as "significant" and rejecting those with "non-significant" results. Considering these concerns, in accordance with ASA recommendations, authors submitting to IJES should refrain from using the terms "statistically significant" or "non-significant" when reporting statistical findings (15).

While interpreting statistics without using these terms may seem atypical, we anticipate that other scholarly sources will ultimately adopt similar policies, as the ASA is encouraging journals to take this step in improving statistical reporting practice (15). In lieu of the binary presentation of findings, subsequent analyses can be utilized to draw conclusions about the data. These include, but are not limited to, consideration of effect sizes, minimal clinically important differences, or analyses employing both frequentist and Bayesian statistical methods.

### Reporting *P*-Values

In light of the elimination of the terms "statistically significant" and "non-significant," authors should report continuous *p*-values. For example, a *p*-value of .072, should be reported as $p = .072$ instead of using the blanket statement of $p > .05$. While this is not a novel recommendation in improving statistical practice, it has recently gained traction (1, 10-11, 15). To be clear, it is still acceptable to indicate $p < .001$ when appropriate. The aim of adopting a requirement for continuous *p*-values is facilitation of improved scientific practice, and encouraging consumers of the information to process the research findings beyond the interpretation of the author(s).

With the ultimate goal of transparency and full disclosure, continuous *p*-values are required for all tested research hypotheses submitted to IJES. Our goal is to avoid the temptation for authors to report only promising findings, which can skew information available to the reader and facilitate bias on a given topic.

When reporting continuous *p*-values, it is important for researchers to be mindful that a *p*-value is not a measure of the magnitude of effect or the impact of the experimental condition (9, 14). In fact, the ASA emphasizes that studies with large sample sizes and precise measures are likely to yield very low *p*-values, while those with smaller samples or less precise measures may see more meaningful changes with *p*-values much closer to .05 (14). Similarly, Gelman and Stern explain that large changes in significance levels can be observed with small, insignificant changes in the underlying values in a practical sense (9).

### Reporting Measures of Effect and Power

Since *p*-values are not indicators of the magnitude of change or difference, the editors are now requiring that authors report effect size alongside each *p*-value, either in the text or in table form. Providing both continuous *p*-values and effect sizes will encourage readers and researchers alike to utilize *p*-values to initiate, but not dictate, scientific thought. Effect size choices should be appropriate for the analyses utilized (2). Moreover, researchers are strongly encouraged to address clinically meaningful and minimum detectable changes for each measured outcome in comparison to observed changes or differences (where appropriate). In addition, we strongly advise the inclusion of a-priori power analyses to allow the reader to understand if the study was adequately powered. If an a-priori power analysis was not conducted due to an exploratory design, post-hoc power must then be included with the reported measures of effect.

## CONCLUSIONS

As researchers and scholarly professionals, we are called to present information transparently to allow others to make informed decisions about our findings. The current research climate has encouraged a simplified practice, where conclusions about practical significance are frequently made based solely on *p*-values at thresholds that cannot be confidently defended. The aforementioned guidelines for statistical reporting are IJES's method of encouraging authors to present findings in a manner that provokes thought and presents a larger picture, regardless of what *appears to be true* based on "statistical significance." The goal of this editorial is to encourage researchers and readers to understand that statistical significance is not synonymous with practical significance. We are hopeful these changes will encourage authors to discuss findings beyond binary significance, addressing the magnitude and meaningfulness of the results.

## REFERENCES

1. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. Am Stat 73(sup1): 262-270, 2019.

2. Borenstein M. Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine. The handbook of research synthesis and meta analysis. pp. 221-237. New York: Russell Sage Foundation; 2009.

3. Boring EG. Mathematical vs. Scientific Significance. Psychol Bull 16(10): 335–338, 1919.

4. Cowles M, Davis C. On the origins of the .05 level of statistical significance. Am Psychol 37(5): 553-558, 1982.

5. Dahiru T. P-value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med 6(1): 21-26, 2008.

6. Edgeworth FY. Methods of Statistics. J Stat Soc London, Jubilee Volume, 181–217, 1885.

7. Fisher RA. Statistical methods for research workers. Oliver and Boyd. Edinburgh, Scotland, 1925.

8. Fisher RA. The arrangement of field experiments. J Ministry Agricul 33: 503-513, 1926.

9. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat 60(4): 328-331, 2006.

10. Hurlbert SH. Levine RA, Utts J. Coup de Grâce for a tough old bull: 'Statistically significant' expires. Am Stat 73(sup1): 352-357, 2019.

11. McShane B, Gal D, Gelman A, Robert C, Tackett J. Abandon statistical significance. Am Stat 73(sup1): 234-245, 2019.

12. Pearson K. On the criterion that a given system of deviations from the probably in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling. London, Edinburgh, and Dublin Philos Mag and J Sci 50: 157-175, 1900.

13. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull 86: 638-641.

14. Wasserstein RL, Lazar NA. ASA statement on statistical significance and p-values. Am Stat 70(2): 129-133, 2016.

15. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05." Am Stat 73(sup1): 1-19. DOI: 10.1080/00031305.2019.1583913, 2019.