INTERNATIONAL JOURNAL OF -EXERCISE SCIENCE-

Editorial

Where is the Power in a Power Analysis?

Kevin D. Dames‡[1], Zoe Climenhaga*[2], Jared Rosenberg†[2]

[1]Biomechanics Laboratory, Kinesiology Department, State University of New York at Cortland, Cortland, NY, USA; [2]Proehl Exercise Physiology Laboratory, Kinesiology Department, State University of New York at Cortland, Cortland, NY, USA

*Denotes student investigator, †Denotes early-career investigator, ‡Denotes established investigator

Abstract

International Journal of Exercise Science 18(2): 1114-1120, 2025. https://doi.org/10.70252/AVGK2063 Strategic planning of research involves predicting the number of replications of the experiment needed to detect an expected effect. The power analysis to determine sample size for the proposed experiment requires known or estimated characteristics of existing distributions. When used well, power analysis reduces the risk of statistical errors, wasted efforts, and temptations to twist ensuing analyses to eke out a 'significant' result after data collections conclude. This editorial highlights some examples of how this process goes awry. Sections are dedicated to the role of researchers, the case for 'pilot' studies, and the critical involvement of reviewers as arbiters of best practices. Throughout, the importance of reporting standard statistical data to support conclusions is identified as the platform for enabling future power analyses. Logical, evidence-based pre-planning of studies and implementing standard statistical reporting increases transparency of research, the likelihood of a study to be cited in the future, and enhances the body of research that exercise scientists collectively build.

Keywords: Alpha and Beta, Type I and Type II error, statistical error

Introduction

Statistics and research methods are two common classes required for undergraduate and graduate degrees. Often, part of the curriculum requires students to critically analyze current studies in their respected fields. For students new to science, who do not yet have the content expertise to spot errors in experimental methodologies or understand the limitations of various instrumentation, the initial critique of a given research study might be that the sample size is too small. The thinking may go like this - a study with 20 participants can't possibly be as conclusive or important as the same study but with 30 participants. The idea that a sample size need only be as large as is necessary to answer the research question comes with mastery. Yet, as their knowledge base develops, students inevitably discover studies with real statistical issues that undermine confidence in the stated conclusions. This could initiate a frank conversation

about the ironic disparity between the standards by which students are graded versus what can pass peer review. The recent editorial by Johnson et al[1] deftly communicates why effect size is critical to report over and above the ubiquitous, but arbitrary, "p <.05". This editorial will build on that foundation, to illustrate why statistical reporting (including effect sizes) is critical to justify reported findings and support new investigations. Lack of transparency in statistical reporting creates a barrier for students planning independent research, a thesis, or dissertation.

While science is built on curiosity, wonder, and exploration, it is also practical and systematic. The mystery of inquiry should not be totally unguided - part of planning new research studies involves estimating potential future effects based on past evidence. The goal of this process is to establish that the effect observed in the experiment truly exists, and is not an artifact of poor research design, statistical missteps, or sampling error.[2] A power analysis involves declaring the anticipated effect size, risk of Type I and II errors, and other situational parameters to support the case of a chosen sample size to test a specific hypothesis.[3] When used appropriately, it enhances the rigor, transparency, and reproducibility of research by ensuring that studies are neither underpowered (leading to inconclusive results) nor excessively overpowered (wasting resources).[2] Power analyses are recommended or required by well-respected journals and publishers (e.g., Journal of Biomechanics, journals within the American Psychological Association, JAMA Neurology [for randomized trials only], and the International Journal of Exercise Science). Software to perform the power analysis is widely available (e.g., G*Power[4], R[5]) and interpretations are well established for various models.[3,6,7] Despite this robust set of resources, issues arise when the process is mishandled, massaged, and/or manipulated. Therefore, the purpose of this paper is to highlight practical issues in power analysis and reiterate the call for transparency in research, aiming to guide students engaged in exercise science research towards a path that strengthens the scientific literature.[8] This paper is not a comprehensive review of any area of research, but will draw on informative examples in some instances. Referencing specific studies confirms the scenario discussed is more than hypothetical but not meant as a censure of the entire work or the authors individually.

What is the purpose of a power analysis?

Researchers might initially answer the question "what is the purpose?", with the obvious – "to justify a sample size for their study." However, the deeper answer might be that pre-planning of their hypothesis prevents p-hacking, Hypothesis After Results Known (HARKing), or other illicit practices that authors may dabble in to reach that coveted $p < a$ threshold.[2] Indeed, rigorous pre-planning of hypotheses or pre-registration of study designs supports replication.[8] Observing an effect that does not truly exist (Type I error) through improper statistical procedures because one expects to have significance contributes to bias in the published literature, which is only later uncovered in systematic reviews, such as that observed in work by Boyer et al.[9] In that paper, the authors identify "a bias towards the publication of studies which report differences between young and older adult gait that agree with the prevalent findings in the literature." Valid, reasonable findings may be blocked from publication, or authors may choose not to submit data that do not align with the response that reviewers currently desire. This issue is often termed the "file drawer problem", where many well-conducted but non-significant studies remain unpublished, skewing the literature toward false

positives. As a result, studies with null or negative results are underreported or excluded. Selective reporting distorts the scientific record, inflates the perceived efficacy or importance of interventions, and undermines the validity of meta-analyses and systematic reviews. Of course, good research practices such as blinding, controls, and reporting of all results (not just the ideal ones) could further increase transparency.[10]

Where are the issues in power analysis from the researcher's side?

As noted above, specific examples from the published literature concretize the noted themes. These papers are not provided as the most egregious cases, nor to censure those authors specifically. To be fair, there are promising and valuable aspects even in flawed studies. An astute reader can appreciate elements which are useful or at least instructive while acknowledging limitations. The most basic example of issues in power analysis from the researcher's side is when details are simply lacking. The reader is left to take the authors' word that the sample size is reasonable rather than seeing the evidence directly. For instance, stating the anticipated effect size without citing any source to substantiate that expectation makes the sample size justification unsupported by data.[11] Citing a source study provides some grounding, but multifaceted issues are present when authors cite a past study in name only. As nearly every study has multiple outcome measures and sometimes multiple parts to the statistical procedures, this leaves the reader wondering which variable was utilized in the analysis and for which statistical model. It may be unclear if the global effect from the ANOVA or the narrower effect in pairwise contrasts was selected. In other cases, authors do not report the quantitative details of the power analysis procedure (e.g., alpha, power level) for a stated design.[12] Failing to report critical details prevents other researchers from replicating the power calculation.

Another mistake could occur when the authors do not adjust for multiple comparisons,[11] leading to an elevated false positive (Type 1 error) rate. This is often found in the literature when multiple *t*-tests are performed instead of an ANOVA with a post-hoc comparison. The form of post-hoc test also controls for family-wise or per-contrast Type I error in unique ways. The Tukey test, for instance, allows all pairwise comparisons and is more liberal, resulting in more frequent significant contrasts. Alternatively, the Dunnett test limits contrasts to only those which compare experimental groups to the control group (eliminating contrasts between all experimental groups). Any 'unnecessary' contrasts conducted reduce statistical power of all desired contrasts. Researchers should not fish for the outcome that produces a desired statistical conclusion retrospectively, but rather pick the approach that suits the research question and plan their sample size to achieve that aim. A related procedural issue that could arise is a circular analysis that is non-independent and facilitates confirmation bias, potentially rendering the findings of the power analysis mute. This can happen when researchers conduct a power analysis using a data set that was already used to estimate effect size for a different hypothesis and/or variable. These conceptual missteps can bias expectations, potentially leading to inflated statistical power and consequently increasing the risk of confirming an effect that was already suggested.

In cases where mathematical procedures and reporting are correct there can still be conceptual missteps. For example, the power analysis strategy does not match the subsequent model(s)

applied, the protocol of the referenced study is unique from the planned study, and/or the population in a source study differs from the one to be sampled. In these cases, the cited effect is not a good indicator of the potential effect. These disjointed purposes can be further compounded when an effect does not scale linearly with changes in the research design or population. For example, a training study lasting 16 weeks may not elicit twice the gain in strength from an 8-week study. The rate of improvement may slow or completely stall. Neither should the adaptations that novices make during 8 weeks of training be used as justification for the potential gains in elite performers (or vice versa). The rapidity of gains in untrained persons is grounded in their greater potential for improvement compared to athletes already near their peak capacity. While the internal validity of a study could be high, the external validity should be considered with an equally critical lens.[13] Recognizing the population to whom experimental treatments might apply, the set or settings in which the effects might reasonably be expected, and signal to noise ratio is important context for interpreting findings.

A power analysis should inform sample size objectively and *a priori*, not be manipulated to conform to outside considerations. There can be a temptation to iteratively try various combinations of data until the result matches a sample size the research team is willing or able to recruit. This can involve inflating the expected effect size or electively lowering the stated power (termed power hacking). Power hacking can potentially undermine the integrity of statistical planning and contributes to irreproducible science. In line with this, other studies have reported an anticipated number of dropouts without specifying where that expectation was derived.[14] A dropout rate of 20% is often used, but without justification, so it is unclear if that percent was chosen arbitrarily or to achieve a desired outcome.[15] Stating an expected dropout rate that is not subsequently observed could overpower the study as the sample analyzed is greater than originally stated. Sprinkling power analysis crumbs in the methods can give the appearance of rigor at a casual read, but upon critical inspection readers may wonder if a convenience sample was defended retroactively during peer-review. Research does not occur in a vacuum, but the context which impacted termination of the study is often not included in the final published article. Even if the research is well-planned according to the availability of time, resources, and willpower, but the power analysis is subsequently manipulated to accommodate these things without regard for the quality of the study, then any power analysis presented in the resulting manuscript is merely perfunctory. Granted, researchers recognize it is more time-consuming and expensive to follow best practices,[8] so some projects may involve intentional sacrifice. Students may appreciate that their thesis or dissertation research needs to meet muster for passing their defense and potential peer review, while the pressing concerns of tuition, graduation timelines, and beginning their career have immediate weight. Meanwhile, their mentor's tenure clocks, grant dollars and deadlines, and semesters are ephemeral.

When are 'pilot' studies truly exploratory?

Pilot studies are small-scale, proof-of-concept investigations that can establish feasibility of a novel treatment before committing to a larger effort. Pilot studies are particularly useful when no pre-existing evidence to inform the proposed project is available. As such, pilot studies are not hypothesis driven.[16] Findings should be discussed as preliminary or exploratory in nature. This groundwork informs subsequent projects which can more rigorously replicate and confirm

the pilot data. A study that claims 'pilot' in the title yet never considers its results in that light sends a mixed signal. Did the reviewers push back on a small sample size, so the authors inserted 'pilot' into the title as a quick alternative to collecting more data? Regardless, this misclassification can lead to confusion, potentially misleading future studies that rely on the data. Sampling or non-sampling errors inherent in a protocol can produce unpredictable effects that do not reflect the true variation in the target population. In addition, satisfying assumptions of parametric tests before applying them can be a concern.[17] Ultimately, the critical role of effect sizes[1] and confidence intervals[18] can inform the researchers, reviewers, and the audience about the statistical (p<.05) or practical significance of the findings. A study may produce statistically significant results of limited practical use. Meanwhile, underpowered (pilot) studies may establish massively practical outcomes that simply do not reach statistical significance due to sample size. Clearer distinction of pre-planned sample sizes vs. convenience samples can help clarify whether the study is truly exploratory.

How should reviewers approach a manuscript's power analysis?

Reviewers are an important part of ensuring published data are of high quality. Reviewers represent a filter that allows studies that meet stringent criteria to pass through. However, study planning confusion and mixed signals are not just limited to the authors, as issues could potentially stem from the peer-review process. The rigor of an acceptable power analysis is not ubiquitous and may drastically differ from journal to journal. A paper rejected from one journal on the basis of insufficient sample size may find acceptance at another without alteration. If a journal does not require certain statistical evidence to support stated conclusions, then subsequent work has no context for replication. For example, a foundational paper in dynamic stability reported no differences in time to stabilization following medial or lateral hops.[19] Consequently, others have cited that experimental protocol but eliminated the medial hop direction, with the expectation that one frontal plane task was sufficient to characterize postural control.[20] Despite other stellar aspects of the original paper that rightly justify its repeated citation and use in designing subsequent experimental protocols, it only provided bar graphs of means by hop direction (forward, backward, sideways) with a table of p values for pairwise comparisons. No F statistic, effect size, or even tabular data of means and standard deviations for each condition (so others could calculate a Cohen's d, for example) is provided as clues for replication and confirmation of the supposed directional symmetry. Details for statistical reporting are readily available[21] and should be referenced when reviewing manuscripts to enhance what the authors initially submit. The reproducibility crisis is broad[8] in part due to (un)intentional under-reporting of statistical details.

In contrast, there is also the case of too much statistical information. Post hoc power analysis is formally requested by some kinesiology journals despite evidence this is not recommended.[16] Type I and II error rates, sample size, and observed effect size are interconnected– relying on sample data to retrospectively confirm one's findings is circular logic.[18] If reviewers request an author include analysis components that are not consistent with best practices, or to remove elements that are essential, then authors find themselves in a tough situation. The authors will have to weigh the choices of acquiescing with the request, despite well-reasoned arguments not to do so, or face rejection. Ultimately, any flaws will be attributed to the intentions of the authors

since the reviewers are not identified unless the journal practices open review. Reviewers and editors should insist the authors justify their sample size adequately and provide necessary statistical data to facilitate replication. Reproducibility is a core tenet of the scientific method, and the referees determining which studies reach public eyes should ensure the projects they approve uphold that promise.

Summary

Transparent statistical reporting and power analysis clarity contribute to reproducible science. Authors simply must report critical statistical data or, as is a growing trend, provide individual data.[22] If scientists, in a perceived 'academic stronghold' hand down knowledge that falls flat on closer inspection, then trust in science by the general public will continue to face challenges.[23] The power analysis process need not be a doom and gloom prospect, an inconvenient hurdle to overcome, or a hastily addressed byline scratched into a manuscript prior to submission; nor should proper statistical reporting be a burden. There are plentiful guides[21] and example data sets with interpretations (https://jasp-stats.org/resources/)[24] to help authors develop their skills. Ultimately, proper planning and reporting can be a boon to citation counts as studies make themselves amenable to future power analysis by other researchers. A well reported power analysis can increase the potential success of a future study, setting the stage for stronger research.

References

1. Johnson SL, Stone WJ, Bunn JA, Lyons TS, Navalta JW. New author guidelines in statistical reporting: Embracing an era beyond p < .05. *Int J Exerc Sci*. 2020;13(1):1-5. https://doi.org/10.70252/HMZN3851

2. Makin TR, De Xivry JJO. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *Elife*. 2019;8:e48175. https://doi.org/10.7554/eLife.48175

3. Cohen J. A Power Primer. *Psychol Bull*. 1992;112(1):155-159. https://doi.org/10.1037/0033-2909.112.1.155

4. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175-191. https://doi.org/10.3758/BF03193146

5. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2024. Accessed December 19, 2024. https://www.r-project.org/

6. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. https://doi.org/10.1016/j.jcm.2016.02.012

7. Qin X. Sample size and power calculations for causal mediation analysis: A Tutorial and Shiny App. *Behav Res Methods*. 2024;56(3):1738-1769. https://doi.org/10.3758/s13428-023-02118-0

8. Baker M, Penny D. Is there a reproducibility crisis? *Nature*. 2016;533:452-454. https://doi.org/10.1038/533452a

9. Boyer KA, Johnson RT, Banks JJ, Jewell C, Hafer JF. Systematic review and meta-analysis of gait mechanics in young and older adults. *Exp Gerontol*. 2017;95:63-70. https://doi.org/10.1016/j.exger.2017.05.005

10. Begley CG. Six red flags for suspect work. *Nature*. 2013;497(7450):433-434. https://doi.org/10.1038/497433a

11. Kalman DS, Feldman S, Krieger DR, Bloomer RJ. Comparison of coconut water and a carbohydrate-electrolyte sport drink on measures of hydration and physical performance in exercise-trained men. *J Int Soc Sports Nutr*. 2012;9. https://doi.org/10.1186/1550-2783-9-1

12. Burrus BM, Moscicki BM, Matthews TD, Paolone VJ. The effect of acute l-carnitine and carbohydrate intake on cycling performance. *Int J Exerc Sci*. 2018;11(2):404-416. https://doi.org/10.70252/JGGV7814

13. Yeh RW, Valsdottir LR, Yeh MW, et al. Parachute use to prevent death and major trauma when jumping from aircraft: Randomized controlled trial. *BMJ (Online)*. 2018;363. https://doi.org/10.1136/bmj.k5094

14. Xue W, Xinlan Z, Xiaoyan Z. Effectiveness of early cardiac rehabilitation in patients with heart valve surgery: a randomized, controlled trial. *J Int Med Res*. 2022;50(7). https://doi.org/10.1177/03000605211044320

15. Zhang L min, Liu Z, Wang J qi, et al. Randomized controlled trial for time-restricted eating in overweight and obese young adults. *iScience*. 2022;25(9). https://doi.org/10.1016/j.isci.2022.104870

16. Levine M, Ensom MHH. Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*. 2001;21(4):405-409. https://doi.org/10.1592/phco.21.5.405.34503

17. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-Test instead of student's t-Test. *Int Rev Soc Psychol*. 2017;30(1):92-101. https://doi.org/10.5334/irsp.82

18. Mair MM, Kattwinkel M, Jakoby O, Hartig F. The minimum detectable difference (MDD) concept for establishing trust in nonsignificant results: A critical review. *Environ Toxicol Chem*. 2020;39(11):2109-2123. https://doi.org/10.1002/etc.4847

19. Liu K, Heise GD. The effect of jump-landing directions on dynamic stability. *J Appl Biomech*. 2013;29(5):634-638. https://doi.org/10.1123/jab.29.5.634

20. Bartlett AS, Mazzone B. Assessment of Jump Type and Leg Dominance on Time to Stabilization (TTS) for Division III Collegiate Athletes: An Exploratory Study. *Adv Orthop Sports Med*. 2022;2022(1).

21. Lang TA, Altman DG. Basic statistical reporting for articles published in Biomedical Journals: The "Statistical analyses and methods in the published literature" or the SAMPL guidelines. *Int J Nurs Stud*. 2015;52(1):5-9. https://doi.org/10.1016/j.ijnurstu.2014.09.006

22. Ferber R, Brett A, Fukuchi RK, Hettinga B, Osis ST. A biomechanical dataset of 1,798 healthy and injured subjects during treadmill walking and running. *Sci Data*. 2024;11(1). https://doi.org/10.1038/s41597-024-04011-7

23. Krause NM, Brossard D, Scheufele DA, Xenos MA, Franke K. Trends - Americans' trust in science and scientists. *Public Opin Q*. 2019;83(4):817-836. https://doi.org/10.1093/poq/nfz041

24. JASP Team. Resources. 2024. Accessed December 19, 2024. https://jasp-stats.org/resources/

Corresponding author: Jared Rosenberg; Jared.rosenberg@cortland.edu