Sample size estimation in sport and exercise science research.

The majority of studies submitted to the *Journal of Sports Sciences* are experimental. Data are collected from a sample of the population and then used to make inferences about that population. A common question in experimental research is therefore "how large should my sample be in order to make precise inferences about the population?". Broadly, there are two approaches to estimating sample size - using power and using precision. If a study uses frequentist hypothesis testing, then it is common to conduct a power calculation to determine how many participants would be required to reject the null hypothesis should an effect of a given size be present. That is, if there's an effect of the treatment (of given size *x*), a power calculation will determine approximately how many participants would be required to detect that effect (of size *x* or larger). Power calculations as conducted in popular software programmes such as G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) typically require inputs for the estimated effect size, alpha, power (1 – ßeta), and the statistical tests to be conducted. All of these inputs are subjective and up to the researcher to decide the most appropriate balance between type 1 error rate (false positive), type 2 error rate (false negative), cost, and time. In contrast, estimating sample size via precision involves estimating how many participants would be required for the confidence or Bayesian credible interval resulting from a statistical analysis to be of a certain width. The implication here is that a narrower confidence or credible interval allows a more precise estimation of where the 'true' population parameter (e.g. mean difference) might be.

To get a sense of the sample sizes and methods used to estimate sample size by studies submitted to the *Journal of Sports Sciences* we randomly selected 120 manuscripts submitted over the previous three years. The data were positively skewed, so the median (median absolute deviation) sample size was 19 (11). Of these 120 manuscripts only 10 % included a formal *a priori* sample size estimation based on power and 1 % estimated sample size using a precision approach. Although the 10 % of manuscripts that did include an *a priori* power calculation identified the effect size to be detected, alpha, and power, 100 % of those manuscripts failed to include full information on the statistical test(s) to be conducted to detect the chosen effect size and 33 % failed to include a convincing rationale for why the given effect size was chosen (e.g. evidence from previous studies).

In order to understand why this is a problem we need to examine the issues with studies that are not adequately powered to detect what could be considered a meaningful effect. As outlined by Brysbaert (2019) and others (Button et al., 2013; Ioannidis, 2005, 2008; Ioannidis, Tarone, & McLaughlin, 2011) the problems with underpowered studies are numerous - the type 2 error rate is increased, if statistically significant effects are detected they will likely overestimate the population effect size (by a considerable amount), statistically significant effects are more likely to be type 1 errors, statistically significant effects are more likely to have low precision in the population estimate, and underpowered studies are less replicable. In regard to overestimating population effect size, the Open Science Collaboration (2015) conducted 100 replications of psychology studies but using high-powered designs and reported that the mean effect size (r = 0.2; ~*d* = 0.4) was approximately half the magnitude of that reported in the original studies. Moreover, Fraley and Vazire (2014) reported that the mean sample size used in psychology studies was 104 participants yet the mean power was only 50% to detect an effect size of *d* = ~0.4 (r = ~0.2). Contrast that with the median sample size of 19 for manuscripts submitted to the *Journal of Sports Sciences* and it's quite likely that we have a problem with underpowered studies in sport and exercise science. Although this is a serious problem, and one we've heard before (Beck, 2013; Heneghan, Perera, Nunan, Mahtani, & Gill, 2012) there are a number of solutions. The obvious solution is to substantially increase the sample size of studies in our field. This would almost certainly increase the power/precision (and quality) of studies and might also lead to a much-needed reduction in the number of studies submitted (the *Journal of Sports Sciences* experienced an increase of 40 % between 2017 and 2019). A reduction in the number of studies would also reduce pressure on over-stretched reviewers.

Although increasing sample size is needed, how sample size is estimated and how data are collected are also important. If studies do indeed conduct an *a priori* sample size estimation they will most likely do so via a power calculation. Yet power analysis serves one purpose – to estimate the sample size required to reject the null hypothesis if indeed there's an effect of a given size. But a power calculation does nothing in regard to estimating the minimum sample size that would ensure a precise estimate of the population parameter. To do this we need to estimate sample size using precision - sometimes called accuracy in

parameter estimation (AIPE) (Kelley, Maxwell, & Rausch, 2003; Kelley & Rausch, 2006; Maxwell, Kelley, & Rausch, 2008). In contrast to the traditional sample size estimation based on power, the AIPE approach bases the sample size estimation on what is required to achieve a certain level of precision in the population parameter effect size. This allows the researcher to identify a width of the confidence interval they would like to achieve (the precision) to estimate the sample size. The width of the confidence interval is proportional to the sample size such that to halve the interval the sample size must increase approximately by a factor of four (Cumming & Calin-Jageman, 2017). The R package MBESS (Kelley, 2019) can be used to estimate sample size using the AIPE approach, as can ESCI software (Cumming & Calin-Jageman, 2017). For example, using the MBESS ss.aipe.smd function, for a standardised mean difference (Cohen's *d*) of 0.4 between two groups, to achieve a 95% confidence interval with a width of 0.6 (0.3 either side of the point estimate) would require a sample size of at least 88. Using the median *Journal of Sports Sciences* sample size of 19 as described earlier a precision (width of the confidence interval) of 1.3 (0.65 either side of the point estimate) would be achieved. This means for d = 0.4 the confidence interval would range from -0.25 (small negative effect) to 1.05 (large positive effect), and therefore such an interval is clearly imprecise. Low precision in the estimated population parameter is not a desired outcome from our research and so outlining *a priori* a desired precision should be encouraged and practiced. Although some argue for a move from using power to using AIPE for sample size estimation (Cumming & Calin-Jageman, 2017; Kelley et al., 2003), the approach still suffers from using a frequentist confidence interval, which is inherently tied to the *p* value and all of its problems (Cohen, 1994; McShane, Gal, Gelman, Robert, & Tackett, 2019; Wasserstein & Lazar, 2016). The confidence interval also contains no distributional information, which means that all values within the interval are equally likely (Kruschke & Liddell, 2018). Actually, the probability of the true population parameter being within the confidence interval is either 1 or 0 because the chosen probability (e.g. 95%) refers to the long-run process of generating the interval, not the interval itself (Barker & R. Schofield, 2008; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Some also argue that because the confidence interval is a theoretical long-run pre-data procedure with a fixed probability (e.g. 95 %), there is no guarantee that a post-data confidence interval will contain the population parameter at all, or have the desired precision (Morey et al., 2016). Most researchers also incorrectly interpret the confidence interval like a Bayesian credible interval (Kruschke & Liddell, 2018), which does contain distributional information and can be used to obtain direct probabilities for the true population parameter (Kruschke, 2013).

Although the AIPE approach allows the researcher to estimate a sample size that will provide adequate precision of the estimated population parameter, it uses a fixed or pre-specified N to do so. This is fine, but what if the effect size is larger or smaller than that estimated *a priori*? An extension of the AIPE approach is to use sequential testing (Kelley, Darku, & Chattopadhyay, 2018; Rouder, 2014). Sequential testing involves collecting data until an *a priori* stopping rule is satisfied. One possible advantage of sequential designs is that sample sizes might be smaller than fixed-N designs yet with the same error rates (Lakens, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Sequential testing can be incorporated into null hypothesis significance testing (NHST) (Kelley et al., 2018; Lakens, 2014), although it has been criticised because only a limited number of interim tests can be performed (Schönbrodt et al., 2017; Wagenmakers, 2007) and Kruschke (2013) contends that it will inevitably lead to a 100% false alarm rate (falsely rejecting the null hypothesis). Alternatively, model comparison (hypothesis testing) or parameter estimation using Bayesian methods avoids such criticisms (Rouder, 2014). That is, when computing Bayes factors (Schönbrodt et al., 2017) or estimating the highest density interval (credible interval) of the posterior distribution, Bayesians are free to monitor the data as often as they wish as it is being collected (Wagenmakers et al., 2018). To help researchers embrace sequential designs when using Bayes factors, Bayes Factor Design Analysis (BFDA) has recently been developed (Schönbrodt & Wagenmakers, 2018; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019). When using a sequential design, BFDA helps researchers determine when data collection should stop once there is strong evidence (as determined by a particular Bayes factor) for either the null hypothesis or the alternative hypothesis. In this scheme, the researcher outlines *a priori* the Bayes factor at which data collection will end (e.g. $BF_{10} > 10$). As the data accumulates the Bayes factor is continuously monitored and once it reaches the set threshold, data collection ceases. Researchers can also set a minimum and maximum N and also determine the probability of obtaining misleading evidence (false positives/negatives). As an example of how to use BFDA, a web-based Shiny app has been developed to allow calculations for an independent-group t-test with directional hypotheses to be performed (Stefan et al., 2019).

As suggested by a number of authors (Cumming, 2014; Kruschke & Liddell, 2018), planning a study based on obtaining a given precision in the population parameter estimate has many advantages over the use of power. Moreover, sequential designs using Bayesian hypothesis testing or parameter estimation also offer a number of advantages over frequentist methods (Rouder, 2014; Schönbrodt & Wagenmakers, 2018). Although we've heard some of these calls before in sport and exercise science (Barker & R. Schofield, 2008; Bernards, Sato, Haff, & Bazyler, 2017), the software required to conduct Bayesian analyses has until recently been inaccessible for many or difficult to use. However, we now have access to a multitude of Bayesian packages in R and also menu-driven software such as JASP (JASP Team, 2020) and SPSS.

Given the discussion so far, the *Journal of Sports Sciences* will now only accept submissions of experimental studies that include a formal *a priori* sample size estimation and rationale. As outlined above, this requirement could be satisfied using (1) a desired level of power to detect a given effect size, (2) the AIPE approach (fixed N or sequential), or (3) a Bayesian sequential design (using Bayes factors and/or parameter estimation), although other methods for power analysis are available (Kruschke, 2013; Weiss, 1997). Whatever the method chosen, authors must report the full range of information required to enable the sample size estimation and/or rationale to be examined and checked by editors, reviewers, and ultimately, by readers. This should include any software used, the exact inputs to calculations, a rationale for those inputs, stopping rules, and the statistical tests used to test a hypothesis or estimate a population parameter. Like any aspect of the method section, those reading the manuscript should be able to replicate your sample size calculations based on the parameters you used and reported and thereby judge if your study is adequately powered and/or precise to answer the research question(s) posed and support the conclusions reached.

We are all probably guilty of conducting underpowered studies, and as such we all have a vested interest in changing the way we plan and conduct our studies. As editors of the *Journal of Sports Sciences* we hope that our new policy outlined here will encourage authors to consider more fully the effect of underpowered studies and how they can change their practice to allow more precise population inferences and ultimately, better science.

Grant Abt
Sports Performance

Colin Boreham
Physical Activity, Health and Exercise

Gareth Davison
Physiology and Nutrition

Robin Jackson
Social and Behavioural Sciences

Alan Nevill
Statistical Advisor

Eric Wallace
Sports Medicine and Biomechanics

A. Mark Williams
Editor-in-Chief

References

Barker, R. J., & R. Schofield, M. (2008). Inference About Magnitudes of Effects. *International Journal of Sports Physiology and Performance*, *3*(4), 547–557. https://doi.org/10.1123/ijspp.3.4.547

Beck, T. W. (2013). The Importance of A Priori Sample Size Estimation in Strength and Conditioning Research. *Journal of Strength and Conditioning Research*, *27*(8), 2323–2337. https://doi.org/10.1519/JSC.0b013e318278eea0

Bernards, J., Sato, K., Haff, G., & Bazyler, C. (2017). Current Research and Statistical Practices in Sport Science and a Need for Change. *Sports*, *5*(4), 87. https://doi.org/10.3390/sports5040087

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, *2*(1), 1–38. https://doi.org/10.5334/joc.72

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

Cohen, J. (1994). The Earth Is Round (p < .05). *American Psychologist*, *49*(12), 997–1003.

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, *25*(1), 7–29.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the New Statistics*. New York: Routledge.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109019

Heneghan, C., Perera, R., Nunan, D., Mahtani, K., & Gill, P. (2012). Forty years of sports performance research and little insight gained. *BMJ*, *345*(jul18 3), e4797–e4797. https://doi.org/10.1136/bmj.e4797

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology*, *22*(4), 450–456. https://doi.org/10.1097/EDE.0b013e31821b506e

JASP Team. (2020). JASP (Version 0.12.2)[Computer software].

Kelley, K. (2019). MBESS: The MBESS R Package. Retrieved from https://cran.r-project.org/package=MBESS

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision. Delineating methods of sample-size planning. *Evaluation & the Health Professions*, *26*(3), 258–287.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. https://doi.org/10.1037/1082-989X.11.4.363

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*(1), 537–563.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

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

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

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. https://doi.org/10.1080/00031305.2016.1154108

Weiss, R. (1997). Bayesian Sample Size Calculations for Hypothesis Testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(2), 185–191.