Kinesiology Review, 2021, 10, 363-368
https://doi.org/10.1123/kr.2020-0067
© 2021 Human Kinetics, Inc.

Human Kinetics
SCHOLARLY REVIEW

The N-Pact Factor, Replication, Power, and Quantitative Research in *Adapted Physical Activity Quarterly*

Jeffrey Martin and Drew Martin

In the current study, a 20-year span of 80 issues of articles ($N = 196$) in *Adapted Physical Activity Quarterly* (*APAQ*) were examined. The authors sought to determine whether quantitative research published in *APAQ*, based on sample size, was underpowered, leading to the potential for false-positive results and findings that may not be reproducible. The median sample size, also known as the N-Pact Factor (NF), for all quantitative research published in *APAQ* was coded for correlational-type, quasi-experimental, and experimental research. The overall median sample size over the 20-year period examined was as follows: correlational type, NF = 112; quasi-experimental, NF = 40; and experimental, NF = 48. Four 5-year blocks were also analyzed to show historical trends. As the authors show, these results suggest that much of the quantitative research published in *APAQ* over the last 20 years was underpowered to detect small to moderate population effect sizes.

Keywords: effect size, false positive, reproducibility, sample size

Attention to the value of reproducing research is increasing in the social sciences (Nosek & Lakens, 2014), exercise and sport psychology (Martin et al., 2019), and disability sport psychology (Atkinson & Martin, 2020). This increased attention started with Ioannidis (2005) and his, at the time, startling claim that most research findings are false. Replicating studies promotes high-quality scientific practice because it helps to confirm or discount the robustness of prior research results and to evaluate the extent to which findings might be unique to a specific sample (Earp & Trafimow, 2015; Martin et al., 2019; Morin, 2016; Mulkay & Gilbert, 1986). If an effect is “real,” the same or a similar effect should be observed in a replication study (Simons, 2014). If a replication fails, it may suggest that a false positive was observed in the initial results (i.e., the conclusion about the hypothesis test was incorrect and involved what is commonly known as a Type I error). Replications can also fail as a result of Type II errors (i.e., false negatives) that can also result from small samples. The goal of all researchers should be to reduce both Type I and Type II errors, and both will occur more frequently with small samples, compared with larger samples.

Despite an increasing call for replication research, however, few researchers conduct replication work (Everett & Earp, 2015; Makel et al., 2012; Maxwell, 2004). Replications may be expensive and time-consuming, take away from a research team’s scholarly agenda, be difficult to publish, and fail to garner as many accolades as the originally conceived work (Coles et al., 2018; Everett & Earp, 2015). If replication research is in short supply, then researchers are particularly dependent on the results of an initial study. However, research findings based on small samples, in comparison with larger samples, are not likely to generalize or replicate because they may be underpowered and produce unreliable findings, such as false positives, false negatives, or effects of smaller or larger magnitude (Button et al., 2013). If studies are underpowered because of small samples, then false positives are likely to reflect large effects. Contrary to common thinking, finding large effects with small samples does not mean that larger samples would have produced similarly large or larger effects. Small samples, relative to large samples, are more likely to miss a true finding, while paradoxically, also being more likely to produce a false positive, as small samples are more susceptible to chance variation. In addition, small samples contribute to significant variations in the p value from study to study, which in turn, can lead to research findings that are irreproducible (Halsey et al., 2015). Finally, researchers have many degrees of freedom in the research decisions they make, leading to questionable research practices (Simmons et al., 2011). Ioannidis (2005) argued that the greater the flexibility researchers have in things like their research design and analyses, the less likely the findings are to be true because researchers can make “negative” or nonsignificant results into “positive” or significant findings. For instance, researchers may find nonsignificant results with an analysis of variance and, in a post hoc attempt to “find” significance, decide to run the same analyses but covary gender. Other researchers may develop their hypotheses to fit the results after seeing the results (i.e., “Hypothesizing After the Results are Known” Kerr, 1998). Finally, John et al. (2012) examined the research practices of over 2,000 psychologists and concluded that questionable research practices were frequent enough to constitute their enactment as the research norm.

It is important to note that the science in crisis (or reproducibility crisis) narrative has been disputed on both empirical and logical grounds. For instance, Fanelli (2018) argued that most “ordinary research,” is not conducted in a theoretical or empirical vacuum, but instead, is grounded in theory and prior research. As a result, researchers are testing predictable hypotheses. Replications are also not the only method for increasing confidence in particular findings and concluding that they are not false positive. For instance, a multitude of studies across various age groups and impairments examining related dependent variables (e.g., fundamental motor skills [FMS], physical activity [PA]) all point toward delayed FMS development and inadequate amounts of PA for children with impairments (e.g., Esposito et al., 2012;

J. Martin is with the Div. of Kinesiology, Health, and Sport Studies, Wayne State University, Detroit, MI, USA. D. Martin is with Bloomfield Hills High School, Bloomfield Hills, MI, USA. J. Martin (aa3975@wayne.edu) is corresponding author.

Lloyd et al., 2014). Inductive reasoning following Mill's Canons (1950) would lead most adapted physical activity (APA) scientists to suggest the above body of work is not indicative of false positives. Jager and Leek (2014) examined over 77,000 abstracts from leading medical journals to determine the false-positive rate, which they determined was 14%. While 14% is not trivial, it is also not suggestive of a crisis. Finally, Klein et al. (2018) conducted a large-scale effort aimed at reproducing the results of 28 classic studies and found evidence for reproducing approximately 50% of the effects, which they reported as consistent with previous work (Camerer et al., 2018).

In terms of false positives, Simmons et al. (2011) have argued that researchers are more likely to incorrectly report evidence for an effect than to correctly discover evidence that an effect is nonexistent. Researchers who report on a true effect (vs. a false positive) do so based on the a priori probability (before the study is done) of it being true, an adequately powered study, and the level used for statistical significance (i.e., the *p* value). All things being equal (e.g., similar measurement error), a study with a large sample size will be more adequately powered than one with a smaller sample size. As a result, research with large samples is more likely to discover small effects that are significant and may be quite meaningful, and less likely to report false positives. If journals in a given discipline routinely publish small-sample research that is underpowered, the field faces a double whammy: small-sample research produces false positives, and the results are not likely to reproduce. Other deleterious results include false negatives (Type II errors). For instance, an intervention study examining the influence of two different (e.g., sport vs. yoga) forms of PA on the FMS of children with developmental delay may find no between-group differences and argue that yoga is just as effective as sport at increasing FMS. However, if the study lacks power, the researchers may be making a Type II error and inaccurately estimating the true population effect.

The sample size of research published in a given field represents or acts as a proxy for whether research is adequately powered, and this is known as the N-Pact Factor (NF: Fraley & Vazire, 2014). Replication research in APA and disability sport may be particularly important to conduct and confirm an effect is true, relative to other disciplines. This is because a brief subjective overview of quantitative research published in the *Adapted Physical Activity Quarterly* (*APAQ*) suggests many studies are based on small samples. Obtaining large samples for parasport or APA research is challenging, as many impairment conditions are rare, making athletes or exercisers with those impairments scarce (Martin, 2017). In other cases, a researcher in adapted physical education may find just a few students with impairments in a physical education class. Relative to abled-bodied youth sport leagues, organized sport opportunities for children and adolescents with disabilities are limited (Martin, 2017). The above factors understandably make conducting research with a large *N* difficult in APA. However, it should be noted that other disciplines and areas of research (e.g., neuroscience, infant research, suicide survivors) also face similar challenges. In addition, an academic culture where the quantity of publications is valued (Ashford, 2013) more than the quality of those publications is also likely to exhibit a plethora of low *N* studies that tend to be underpowered. In addition, with a few exceptions, replication research in APA is rare (Atkinson & Martin, 2020; Martin et al., 2015).

The NF was introduced by Fraley and Vazire (2014), who conducted a study on sample sizes in *Personality and Social Psychology*. Schweizer and Furley (2016) also examined the NF in exercise and sport psychology research. Both sets of researchers found that sample sizes in those disciplines tended to be small, with research being underpowered and unable to detect small to moderate effect sizes. The earlier suggestion, that research published in *APAQ* is based on small samples resulting in underpowered research, however, is a subjective assessment. Hence, to address this shortcoming, the first purpose of this research report was to examine the median sample size of quantitative research published in *APAQ* over the last 20 years. The second purpose was to look at any potential trends in the NF over the last 20 years by examining the data in four sequential 5-year blocks. Finally, a third critical goal was to determine the power of that same published research to detect various population effect sizes.

Method

A total of 20 years (80 issues) of *APAQ* were examined (2000–2019) in the current study. All published papers were categorically coded as an editorial, viewpoint, review, meta-analysis, psychometric research (i.e., the major purpose was examining validity or reliability), qualitative, and quantitative research. The focus of the current analyses was on the 196 quantitative research studies published in *APAQ*. The quantitative research was then coded as examining relationships (i.e., correlations, multiple regression, structural equation modeling) or group differences (i.e., *t* tests, analysis of variance). The latter category was coded in two ways: differences between nonrandom assigned groups if group membership did not result from random assignment or, alternatively, randomly assigned group differences if group membership was the result of random assignment. This coding scheme was grounded in prior work by Schweizer and Furley (2016) and Fraley and Vazire (2014) to allow for intradisciplinary comparisons. The first author coded the articles twice, 8 weeks apart, and compared the results. In addition, a random sample of 20% of the papers was recoded a third time by each author, with 96% agreement. The authors discussed and resolved differences. Difficulties in coding were often the result of statistical tests being conducted that reflected both correlational and difference analyses. These differences were resolved by determining what the major purpose of the study was, as stated by the authors.

The overwhelming number ($N = 137$) of quantitative research studies published in *APAQ* was quasi-experimental research at an average of 1.71 per issue. A common quasi-experimental study design was comparing a group of children with impairments to typically developing children on various dependent variables (e.g., motor skills, PA). Correlation and correlational types (e.g., multiple regression) of studies examining relationships were next ($N = 50$), followed by mixed (within and between differences) experimental research ($N = 9$). The median or the NF was used because the mean is too susceptible to studies with very large *N*s. For instance, *APAQ* has published a few studies based on national databases where tens of thousands of participants are included (Haegele et al., 2019). For power calculations, we followed the supplemental directions found in Schweizer and Furley (2016) based on Faul et al. (2009) and, in all cases, used the post hoc power function with a two-tailed test. The exact family test was used for correlational studies, and the *t* test for the quasi and experimental studies. In general, for the current study, we followed much of the rationale and procedures of Fraley and Vazire (2014) and Schweizer and Furley.

Results and Discussion

The first goal of this study was to examine the NF, or median sample size, of research published in *APAQ* (see Table 1). Table 1 also includes the interquartile ranges and provides more detailed data about the sample size distribution based on the lowest and highest quartile. The overall median sample size over the 20-year period examined was as follows: correlational types, NF = 112; quasi-experimental, NF = 40; and experimental, NF = 48. For correlational and quasi-experimental, the largest NF occurred during the last 5-year block. However, the correlational results reflect a U-shaped curve, with the highest NFs in the first and last 5-year blocks and the lowest NFs in the middle two 5-year blocks. For the quasi-experimental row, the trend is quite stable, with three 5-year blocks exhibiting NFs of 45, 46, and 47. For the experimental studies, the two most recent 5-year blocks had the highest NF, but these results are based on very few studies. The above pattern of results suggests some very tentative optimism and that the value of adequate sample sizes may be reaching *APAQ* authors and reviewers. In a similar study by Schweizer and Furley (2016), examining research in sport and exercise psychology journals from 2009 to 2013, the results were as follows: correlational types, NF = 221; quasi-experimental, NF = 91; and experimental, NF = 40. The larger NF for correlational types (221 vs. 112) and quasi-experimental (91 vs. 40), in sport and exercise psychology journals might most reasonably be attributed to the sample size constraints associated with conducting research with individuals with disabilities relative to research with able-bodied individuals. It is also important to note that, over the course of 20 years, *APAQ* has only published nine studies where completely randomized designs were employed. Randomized designs are considered the gold standard of quantitative research because of their ability to support cause-and-effect inferences. Hence, their scarcity in *APAQ* publications suggests one area for improvement.

Another driving factor of the current study was to determine whether *APAQ* research is adequately powered to detect the most common effect sizes, as shown in Tables 2 and 3. An example of how to interpret Table 2 is provided next. In Row 1 of Table 2, all the results in the five columns pertain to correlational studies from 2000 to 2019, where the median sample size is 112. In Column 1,

Table 1 The Median (NF) Sample Size per 5-Year Blocks and 20 Years

Type of study	2000–2004	2005–2009	2010–2014	2015–2019	2000–2019
Correlational	130 (96, 263) *n* = 9	103 (66, 130) *n* = 11	49 (31, 155) *n* = 16	167 (49, 345) *n* = 14	112 (48, 213) *N* = 50
Quasi-experimental	46 (24, 105) *n* = 51	30 (20, 77) *n* = 35	45 (20, 60) *n* = 19	47 (16, 97) *n* = 32	40 (20, 83) *N* = 137
Experimental	32 (22, X) *n* = 3	48 (46, X) *n* = 3	131 (28, X) *n* = 2	114 (X, X) *n* = 1	48 (30, 114) *N* = 9

Note. The numbers within parentheses are interquartile ranges. The *n* refers to the number of studies published in that category for that time period. NF = N-Pact Factor.

Table 2 Statistical Power to Detect Various Population Effect Sizes Across Study Categories

	r = .1 (*d* = 0.2)	*r* = .2 (*d* = 0.41)	*r* = .3 (*d* = 0.63)	*r* = .4 (*d* = 0.87)	*r* = .5 (*d* = 1.15)
Years 2000–2019					
Correlational (NF = 112)	.18	.57	.90	.99	.99
Quasi-experimental (NF = 40)	.09	.24	.49	.76	.94
Experimental (NF = 48)	.10	.28	.57	.84	.97
Years 2015–2019					
Correlational (NF = 167)	.25	.74	.98	.99	.99
Quasi-experimental (NF = 47)	.10	.28	.57	.84	.97
Experimental (NF = 114)	.18	.58	.92	.99	.99

Note. Correlational (*r*) effect size as the reference: .1 = small, .3 = medium, .5 = large, and mean difference (*d*) equivalent. NF = median sample size.

Table 3 Statistical Power to Detect Various Population Effect Sizes Across Study Categories

	d = 0.2 (*r* = .1)	*d* = 0.35 (*r* = .17)	*d* = 0.5 (*r* = .24)	*d* = 0.65 (*r* = .31)	*d* = 0.8 (*r* = .37)
Years 2000–2019					
Correlational (NF = 112)	.18	.45	.75	.93	.99
Quasi-experimental (NF = 40)	.09	.19	.34	.52	.69
Experimental (NF = 48)	.10	.22	.39	.59	.77
Years 2015–2019					
Correlational (NF = 167)	.25	.61	.89	.99	.99
Quasi-experimental (NF = 47)	.10	.21	.38	.58	.75
Experimental (NF = 114)	.18	.46	.75	.93	.99

Note. Mean difference (*d*) effect size as the reference: 0.2 = small, 0.5 = medium, 0.8 = large, and correlation (*r*) equivalent. NF = median sample size.

the results only refer to the ability of such a study to detect a significant correlation of .1 or a mean difference of 0.2. The finding, of 0.18, indicates that such a study only has an 18% chance of determining that a small effect size ($r = .1$, $d = 0.2$) was significant. In contrast, the last column for the same row indicates that there was a 99% chance of detecting a large effect size ($r = .5$, $d = 1.15$). In general, large effect sizes are more meaningful than small effect sizes.

However, given the research question and context, small effect sizes can be very meaningful, and large effect sizes trivial. For instance, the influence of skill in producing a single successful at-bat for professional baseball players is one third of 1% (Abelson, 1985). This paradoxical and counterintuitive finding becomes understandable when considering that the influence of skill across a full season or of multiple skilled players batting one after another in one inning was not taken into account. In terms of exercise behavior, researchers have found a small positive effect of exercise on both memory ($d = 0.24$) and executive function ($d = 0.27$) in older adults (Sanders et al., 2019). The translation of what a small effect size means for everyday behavior is rarely addressed. However, if such a small effect translates into remembering where the hotel is in a large foreign city when out on a run, or where the car is in a shopping mall car park at night, most people are likely to judge such a small effect as quite meaningful.

As the above sport and exercise examples illustrate, effect sizes should always be interpreted within the context of the research question. However, in general, Cohen (1988) suggested that $d =$ 0.2 is a small effect, $d = 0.5$ is medium, and $d = 0.8$ is large. As can be seen from Table 2, the ability to detect a small effect size is quite low, ranging from 9% to 18%. The ability to detect an effect size closer to moderate ($d = 0.41$) is better, at 24–57%. Given that *APAQ* publishes a lot of quasi-experimental research (e.g., comparing two or more nonrandomized groups on various dependent variables), the 24% ability to detect a true effect for $d = 0.41$, or a correlation of only .20, should be regarded with some disquiet.

Few journal editors publish research with null findings (Fanelli, 2012), often called the “file drawer” problem (Rosenthal, 1979). As a result, if we assume that there is a tendency for *APAQ*, like most journals, to publish research that involves significant findings, then most of that research will involve medium and large effect sizes. As noted earlier, finding large effects might initially seem impressive if they are being found with small (e.g., $N = 40$) samples. Recall that finding large effects with small samples does not mean a similar study with a larger sample will produce similarly large or larger effects. Small samples, relative to large samples (all else being equal), are more likely to produce both false positives and false negatives due to the influence of just a few participants’ data on the results. For example, one participant’s score exerts half the influence on the results in a study with an N of 2. In contrast, one participant in a sample of 100 has a 1% influence on the results. The above example and rationale are why statistical tests on samples of four to six people are generally frowned upon by journal reviewers and statisticians unless the statistics are simple descriptives, such as percentages or means.

Limitations, Conclusions, and Research Implications

In the current study, we examined one element within the broader research enterprise: the NF. Clearly, there are many other elements of research that are important considerations, such as the research questions, whether theory is used, measurement error, contribution to advancing the literature, and practical implications. The single focus on NF is not intended to diminish the value of other important research considerations. Power and sample size are strongly related, but not equivalent (Maxwell, 2004). Hence, while we have strongly emphasized the importance of sample size in producing an adequately powered study, there are two caveats to this point. First, there are other ways to increase power, in addition to obtaining more participants, as noted by various authors (Hansen & Collins, 1994; Lazic, 2018; McClelland, 2000). Avoiding redundant (highly correlated) predictors in a regression equation and using measures with minimal measurement error (i.e., scales that produce scores considered valid and reliable) are just two of these ways (Lilienfeld & Strother, 2020). Using state of the art techniques to replace missing data to avoid deleting participants is strongly recommended (Little et al., 2014), as well as conducting planned missing data research designs to promote better quality data (Moore et al., 2020).

Second, it should be acknowledged that increasing the sample size can make conducting a study more expensive and difficult and can have ethical implications. People with disabilities frequently experience chronic pain and might experience pain, fatigue, and even injury if engaging in lengthy and/or intense exercise intervention studies (Martin, 2017). As Bacchetti et al. (2005) pointed out, the inconvenience and time spent in research by participants often outweigh any benefits they personally receive. Many individuals with disabilities have reduced incomes, lack transportation and social support, and spend more time doing simple activities of daily living, compared with able-bodied people. This means the net burden on them may be even greater than the burden on a person without a disability. For example, participating in a research study may be an all-day time investment for a person with a disability and leave them quite fatigued. In contrast, for a person without a disability, it may represent a much smaller time commitment and investment of energy. Such a burden is clearly compounded with increased sample sizes.

Given that we report on group data, there are likely many individual studies within the 20-year period of *APAQ* that are likely not underpowered. Fraley and Vazire (2014) suggested that a focus on large sample sizes should not be done at the expense of multimethod, ecologically sound, or longitudinal studies because of the difficulty of obtaining adequate sample sizes. The same rationale is even more applicable to APA research. Relative to other disciplines, APA researchers face unique challenges to obtaining adequate sample sizes. As a long-time researcher in disability sport and exercise psychology, the first author is intimately familiar with the challenges of obtaining large samples. For instance, over 20 years ago, he once spent 3 years traveling to and attending the same three elite-level road races in order to obtain a sufficient sample of elite wheelchair marathoners.

The purpose of the current study and attendant commentary was to highlight the importance of sample size to researchers and reviewers. In turn, as they design research studies, evaluate research, and move the field forward, researchers will have to make individual choices about the importance of sample size. However, reviewers appear to be more alert to the negative implications of small sample sizes, suggesting such considerations may increasingly play a role in reviewing decisions. In addition, as the *APAQ* rejection rate appears to be increasing, publishing in *APAQ* is likely to become more competitive. In brief, authors may increasingly be asked to justify their sample size, as publication decisions may become more contingent on sample size and power. Currently, *APAQ* requires a power analysis to support authors' sample-size decisions. Authors may also consider estimated sample size with an accuracy in parameter estimation, which provides a confidence interval estimate. The width of the confidence interval is related to the sample size and helps researchers determine a minimum sample size to ensure a precise estimate of the population parameter (Maxwell et al., 2008). Furthermore, the accuracy in parameter estimation accounts for multiple comparisons, which APA researchers frequently engage in (Pan & Kupper, 1999).

Aside from the obvious recommendation to increase the sample size of APA research, researchers should consider the following related recommendations. Researchers have developed a number of effect size adjustments, and we briefly discussed just a few as examples of what APA researchers may consider. Unadjusted effect sizes are specific to the sample that generated them, whereas, adjusted effect sizes reflect the population effect size (Thompson, 2006). When research is conducted with small samples, the adjusted effect size should be used because the unadjusted effect size is biased upwards. For instance, using Ezekiel's (1930) formula, for a sample of 20, an unadjusted effect size of 0.50 would result in an adjusted effect size of 0.16 (Ivarsson et al., 2013). In contrast, with a sample of 80, the adjusted effect size (0.44) is much closer to the unadjusted effect size of 0.50. Although it is beyond the scope of this paper to provide detailed explanations of other ways to interpret effect sizes, we offer two that Ivarsson et al. (2013) suggested, which are applicable to APA research, given its heavy focus on quasi-experimental research, as seen in *APAQ* publications. First, one effect size interpretation is the number needed to treat (NNT) and is typically used for dichotomous dependent variables (Martinez-Gutierrez et al., 2019). Many APA researchers have used the Test of Gross Motor Development to determine if a motor skill or PA intervention was successful in developing a child's FMS (Ketcheson et al., 2021).

The dichotomous scoring procedure (i.e., 1 if the skill is present and 0 if it is not present) makes it suitable for determining the NNT. Children in the intervention who do not improve (a percentage) are subtracted from the percentage who do not improve in the control group, resulting in the risk difference, and the NNT is 1/risk difference. For instance, if 2/20 (10%) do not improve in the intervention group, and 18/20 (90%) in the control group do not improve, the risk difference is 90-10, or 0.80, and the NNT is $1/0.80 = 1.25$, indicating that four out of five children improved. The value of having four out of five children improve can then be weighed against the cost (e.g., time, money, etc.). Second, the probability of superiority index (Fritz et al., 2012) allows researchers to determine the percentage of participants, if chosen randomly from a successful intervention group (e.g., a group with a higher mean Test of Gross Motor Development score), who would score higher than a randomly chosen participant from the control group (with a lower mean Test of Gross Motor Development score).

Summary

It appears that *APAQ* published research is often underpowered, by virtue of small samples, to detect small to moderate effect sizes. As a result, it is possible that false-positive (and false-negative) findings are being published that may be difficult to replicate. While recognizing the difficulty of increasing sample sizes, APA researchers are, nonetheless, urged to make stronger efforts to increase their sample sizes, justify their sample sizes with an accurate power analysis or accuracy in parameter estimation, and do more to interpret their effect sizes.

References

Abelson, R.P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*(1), 129–133. https://doi.org/10.1037/0033-2909.97.1.129

Ashford, S.J. (2013). Having scholarly impact: The art of hitting academic home runs. *Academy of Management Learning & Education, 12*(4), 623–633. https://doi.org/10.5465/amle.2013.0090

Atkinson, F., & Martin, J. (2020). Gritty, hardy, resilient, and socially supported: A replication study. *Disability and Health Journal, 13*(1), 100839. PubMed ID: 31519505 https://doi.org/10.1016/j.dhjo.2019.100839

Bacchetti, P., Wolf, L.E., Segal, M.R., & McCulloch, C.E. (2005). Ethics and sample size. *American Journal of Epidemiology, 161*(2), 105–110. PubMed ID: 15632258 https://doi.org/10.1093/aje/kwi014

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., & Munafò, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. PubMed ID: 23571845 https://doi.org/10.1038/nrn3475

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637–644. PubMed ID: 31346273 https://doi.org/10.1038/s41562-018-0399-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Coles, N.A, Tiokhin, L., Scheel, A.M., Isager, P.M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences, 41,* e124. https://doi.org/10.1017/S0140525X18000596

Earp, B.D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6,* 621–632. PubMed ID: 26042061 https://doi.org/10.3389/fpsyg.2015.00621

Esposito, P.E., MacDonald, M., Hornyak, J.E., Ulrich, D.A. (2012). Physical activity patterns of youth with Down syndrome. *Intellectual and Developmental Disabilities, 50*(2), 109–119. PubMed ID: 22642965 https://doi.org/10.1352/1934-9556-50.2.109

Everett, J.A.C., & Earp, B.D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6,* 1152. PubMed ID: 26300832 https://doi.org/10.3389/fpsyg.2015.01152

Ezekiel, M. (1930). *Methods of correlational analysis*. Wiley.

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

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences, 115*(11), 2628–2631. https://doi.org/10.1073/pnas.1708272114

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. PubMed ID: 19897823 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), e109019. PubMed ID: 25296159 https://doi.org/10.1371/journal.pone.0109019

Fritz, C.O., Morris, P.E., & Richler, J.J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. https://doi.org/10.1037/a0024338

Haegele, J.A., Aigner, C.J., & Healy, S. (2019). Prevalence of meeting physical activity, screen-time, and sleep guidelines among children and adolescents with and without visual impairments in the United States. *Adapted Physical Activity Quarterly, 36*(3), 399–405. PubMed ID: 31155913 https://doi.org/10.1123/apaq.2018-0130

Halsey, L.G., Curran-Everett, D., Vowler, S.L., & Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179–185. PubMed ID: 25719825 https://doi.org/10.1038/nmeth.3288

Hansen, W.B., & Collins, L.M. (1994). Seven ways to increase power without increasing N. *NIDA Research Monograph, 142,* 184–195. PubMed ID: 9243537

Ioannidis J.P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. PubMed ID: 16060722 https://doi.org/10.1371/journal.pmed.0020124

Ivarsson, A., Andersen, M.B., Johnson, U., & Lindwall, M. (2013). To adjust or not adjust: Nonparametric effect sizes, confidence intervals, and real-world meaning. *Psychology of Sport and Exercise, 14*(1), 97–102. https://doi.org/10.1016/j.psychsport.2012.07.007

Jager, L.R., & Leek, J.T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics, 15*(1), 1–12. PubMed ID: 24068246 https://doi.org/10.1093/biostatistics/kxt007

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. PubMed ID: 22508865 https://doi.org/10.1177/0956797611430953

Kerr, N.L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. PubMed ID: 15647155 https://doi.org/10.1207/s15327957pspr0203_4

Ketcheson, L.R., Centeio, E.E., Snapp, E.E., McKown, H.B., & Martin, J.J. (2021). Physical activity and motor skill outcomes of a 10-week intervention for children with intellectual and developmental disabilities ages 4–13: A pilot study. *Disability and Health Journal, 14*(1), 100952. PubMed ID: 32624452 https://doi.org/10.1016/j.dhjo.2020.100952

Klein, R.A., Vianello, M., Hasselman, F., Adams, B.G., Adams, R.B., Alper, S., . . . Nosek, B.A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Lazic, S.E. (2018). Four simple ways to increase power without increasing the sample size. *Laboratory Animals, 52*(6), 621–629. PubMed ID: 29629616 https://doi.org/10.1177/0023677218767478

Lilienfeld, S.O., & Strother, A.N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology, 61*(4), 281–288 https://doi.org/10.1037/cap0000236

Little, T.D., Jorgensen, T.D., Lang, K.M., & Moore, E.W.G. (2014). On the joys of missing data. *Journal of Pediatric Psychology, 39*(2), 151–162. PubMed ID: 23836191 https://doi.org/10.1093/jpepsy/jst048

Lloyd, M., Saunders, T.J., Bremer, E., & Tremblay, M.S. (2014). Long-term importance of fundamental motor skills: A 20-year follow-up study. *Adapted Physical Activity Quarterly, 31*(1), 67–78. PubMed ID: 24385442 https://doi.org/10.1123/apaq.2013-0048

Makel, M.C., Plucker, J.A., & Hegarty, B. (2012). Replications in psychology research how often do they really occur? *Perspectives in Psychological Science, 7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Martin, J., Beasley, V., & Guerrero, M. (2019). Sport psychology research: Proper standards and limitations. In M.H. Anshel (Ed.), *Handbook of sport and exercise psychology* (pp. 17–40). American Psychological Association.

Martin, J.J. (2017). *Handbook of disability sport and exercise psychology*. Oxford University Press.

Martin, J.J., Byrd, B., Watts, M.L., & Dent, M. (2015). Gritty, hardy, and resilient: Predictors of sport engagement and life satisfaction in wheelchair basketball players. *Journal of Clinical Sport Psychology, 9*(4), 345–359. https://doi.org/10.1123/jcsp.2015-0015

Martinez-Gutierrez, J.C., Leslie-Mazwi, T., Chandra, R.V., Ong, K.L., Nogueira, R.G., Goyal, M., . . . Hirsch, J.A. (2019). Number needed to treat: A primer for neurointerventionalists. *Interventional Neuroradiology, 25*(6), 613–618. PubMed ID: 31248312 https://doi.org/10.1177/1591019919858733

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163. PubMed ID: 15137886 https://doi.org/10.1037/1082-989X.9.2.147

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. PubMed ID: 17937603 https://doi.org/10.1146/annurev.psych.59.103006.093735

McClelland, G.H. (2000). Increasing statistical power without increasing sample size. *American Psychologist, 55*(8), 963–964. https://doi.org/10.1037/0003-066X.55.8.963

Mill, J.S. (1950). *Philosophy of scientific method*. Hafner Publishing Co.

Moore, E.W.G., Lang, K.M., & Grandfield, E.M. (2020). Maximizing data quality and shortening survey time: Three-form planned missing data survey design. *Psychology of Sport and Exercise, 51,* 101701. https://doi.org/10.1016/j.psychsport.2020.101701

Morin, K.H. (2016). Replication: Needed now more than ever. *Journal of Nursing Education, 55*(8), 423–424. https://doi.org/10.3928/01484834-20160715-01

Mulkay, M., & Gilbert, G.N. (1986). Replication and mere replication. *Philosophy of the Social Sciences, 16*(1), 21–37. https://doi.org/10.1177/004839318601600102

Nosek, B.A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology, 45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Pan, Z., & Kupper, L.L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine, 18*(12), 1475–1488

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sanders, L.M., Hortobágyi, T., la Bastide-van Gemert, S., van der Zee, E.A., & van Heuvelen, M.J. (2019). Dose–response relationship between exercise and cognitive function in older adults with and without cognitive impairment: A systematic review and meta-analysis. *PLoS One, 14*(1), e0210036. PubMed ID: 30629631 https://doi.org/10.1371/journal.pone.0210036

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise, 23,* 114–122. https://doi.org/10.1016/j.psychsport.2015.11.005

Simmons, J., Nelson, L., & Simonsohn, U. (2011). Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. PubMed ID: 22006061 https://doi.org/10.1177/0956797611417632

Simons, D.J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80. PubMed ID: 26173243 https://doi.org/10.1177/1745691613514755

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. Guilford Press.

Copyright of Kinesiology Review is the property of Human Kinetics Publishers, Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.