Psychology of Sport and Exercise 23 (2016) 114–122

Contents lists available at ScienceDirect

Psychology of Sport and Exercise

journal homepage: www.elsevier.com/locate/psychsport

Reproducible research in sport and exercise psychology: The role of sample sizes

Geoffrey Schweizer [a, *], Philip Furley [b]

[a] University of Heidelberg, Department of Sport and Exercise Psychology, Im Neuenheimer Feld 720, D-69120 Heidelberg, Germany
[b] German Sport University Cologne, Institute of Cognitive and Team/Racket Sport Research, Am Sportpark Müngersdorf 6, D-50933 Köln, Germany

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form
4 November 2015
Accepted 20 November 2015
Available online 2 December 2015

Keywords:
Replicability
Power
False positive
Effect size
Research methods

ABSTRACT

Objectives: We aim to introduce the discussion on the crisis of confidence to sport and exercise psychology. We focus on an important aspect of this debate, the impact of sample sizes, by assessing sample sizes within sport and exercise psychology. Researchers have argued that publications in psychological research contain numerous false-positive findings and inflated effect sizes due to small sample sizes.
Method: We analyse the four leading journals in sport and exercise psychology regarding sample sizes of all quantitative studies published in these journals between 2009 and 2013. Subsequently, we conduct power analyses.
Results: A substantial proportion of published studies does not have sufficient power to detect effect sizes typical for psychological research. Sample sizes and power vary between research designs. Although many correlational studies have adequate sample sizes, experimental studies are often underpowered to detect small-to-medium effects.
Conclusions: As sample sizes are small, research in sport and exercise psychology may suffer from false-positive results and inflated effect sizes, while at the same time failing to detect meaningful small effects. Larger sample sizes are warranted, particularly in experimental studies.

> "At its core, this crisis is about the justification of knowledge. Researchers in the social and behavioral social sciences make claims that are simply unsupported by the methods they use."

With this quote, Hoekstra, Morey, Rouder, and Wagenmakers (2014, p. 1162) refer to the so-called crisis of confidence that has been an alarming topic in psychological research during the past years. While the crisis of confidence has led to numerous publications (e.g., Fraley & Vazire, 2014; Open Science Collaboration, 2012, in press), several special issues (e.g., Nosek & Lakens, 2014; Pashler & Wagenmakers, 2012; Spellman, 2012) and even the emergence of new organizations or projects dedicated to improving the quality of psychological research (Open Science Collaboration, 2012; in press), the field of sport and exercise psychology has been relatively unaffected by this movement. This does not mean that scholarly contributions calling for high standard research methodology have been absent within this field: see for example Myers, Ahn, and Jin (2011) on a special instance of estimating sample sizes and power, or Zhu (2012) on Null Hypothesis Significance Testing, or reliability (Zhu, 2013). However, an alarming series of failed replications (Open Science Collaboration, 2015) has cast justified, new doubt on common research and publishing practices. In a first step of remedying the main problems causing the failed replications, the field of psychology has provided evidence for some of the most important shortcomings within psychology. Following this approach the present manuscript has the specific aim of introducing this discussion to sport and exercise psychology, and more importantly of reviewing the state of the field regarding a crucial aspect of the discussion, namely the adequacy of sample sizes.

1. The current debate: a crisis of confidence

The so-called *crisis of confidence*, sometimes also called *replication crisis* centers around the observation that it has been surprisingly hard to replicate well-known psychological results (e.g., Ebersole et al., 2015; Nosek & Lakens, 2014; Open Science Collaboration, 2015). Furthermore, it has been noted that replications have been underappreciated by the scientific community and publication outlets and therefore replication attempts are hard to find in the psychological literature (Makel, Plucker, &

* Corresponding author.
E-mail address: geoffrey.schweizer@issw.uni-heidelberg.de (G. Schweizer).

http://dx.doi.org/10.1016/j.psychsport.2015.11.005
1469-0292/

Hegarty, 2012; Open Science Collaboration, 2012). At the same time, researchers have cast doubt on the quality of the research methods psychologists commonly use (Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). For example, several papers have compellingly argued that psychological research suffers from an inflated rate of false-positives or Type 1 errors, and they have identified factors contributing to the inflation of false-positives (Simmons et al., 2011). A false-positive means that researchers report finding an effect that does not exist in the real world. In other words, researchers reject the null hypothesis although it would have been true (Simmons et al., 2011). Although most researchers may be aware of some risk of reporting a false-positive, the problem is that many researchers do not seem to be aware of the magnitude of this risk and of what factors contribute to it. The actual risk of reporting a false-positive depends on several factors, and given certain constellations of factors it may well exceed 50%, as Simmons and colleagues convincingly showed (2011).

> … we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤.05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not … (Simmons et al., 2011, p. 1359)

This number is neither a new one nor confined to psychological research: As early as in 2005, Ioannidis warned that in biomedical research about 50% of all reported research findings may be wrong (2005). But why is that? The inflation of false-positive rates is usually explained by the interplay of several factors (Ioannidis, 2005; Simmons et al., 2011). The most prominent among them are researcher degrees of freedom, the file-drawer effect, and sample sizes.

Researcher degrees of freedom refer to choices researchers can make when collecting, analyzing, and reporting their data (Simmons et al., 2011). For example, researchers may sequentially increase their sample size, choose between several covariates or between several dependent variables, or they may selectively drop experimental conditions. When researchers employ any of these behaviours or combinations of them, and the null hypothesis is actually true, then the probability that their study will yield a false-positive result increases. In turn, when significant results have a higher likelihood of being published, the proportion of false-positive results in relation to correct-positive results gets higher (Simmons et al., 2011). When systematic replication attempts are scarce, these factors lead to an ever increasing proportion of false positive results, that may go undetected for a long time. Whereas the contribution of researcher degrees of freedom and selective publishing of significant results is intuitively appealing, the role of sample sizes is far less intuitive to understand. This is despite the fact that sample sizes may play the most important role, both in understanding the crisis of confidence and in finding remedies against this crisis.

2. Adequate sample sizes

Traditionally, sample sizes have been discussed within the power framework (Cohen, 1962, 1992; Sedlmeier & Gigerenzer, 1989). Power is typically defined in terms of Null Hypothesis Significance Testing (NHST). In NHST, power is defined as the probability of getting a significant result in a study, given the null hypothesis is wrong. In other words, the power of a test is the probability of correctly rejecting the null hypothesis (Button et al., 2013; Gelman & Carlin, 2014). Within this framework, two kinds of errors can be made. A Type 1 error means rejecting the null hypothesis although it is true, whereas a Type 2 error means accepting the null hypothesis although it is false. A Type 1 error is sometimes also called a false-positive and a Type 2 error a miss (Fraley & Vazire, 2014).

Some researchers deliberately refrain from referring to the concepts power, Type 1 error, and Type 2 error because these are only relevant from a NHST point of view which has received increasing criticism lately (Cumming, 2012; 2014; Gelman & Carlin, 2014). Alternatively, Gelman and Carlin (2014) suggest using the terms Type M error and Type S error. A Type M error is an error of magnitude that means it refers to the degree to which the estimate of an effect in a study deviates from the real effect. A Type S error is a sign error that means it occurs when the estimate of an effect in a study has the wrong sign. The main difference between the former and the latter concepts is that the former (i.e., Type 1 error and Type 2 error) refer to a binary concept, i.e. a hypothesis is either rejected or it is not. The latter (i.e., Type S error and Type M error) assume that we want to estimate an effect and that we can quantify the degree to which the estimate is off. It is important to note that the relevance of sample sizes is not confined to studies using NHST. In recent years, opposition to NHST (see also Zhu, 2012 within the field of sport science) has increased, and many researchers strongly suggest abolishing NHST for most study purposes (Cohen, 1994; Cumming, 2012; Wagenmakers, 2007). The journal Basic and Applied Social Psychology has even banned NHST in research articles (Trafimow & Marks, 2015). Suggested alternatives are the estimation of confidence intervals around effect sizes (Cumming, 2012; 2014) and Bayesian methods (Lee & Wagenmakers, 2005; Wagenmakers, 2007; Wetzels et al., 2011). Both parameter estimation and Bayesian methods require adequate sample sizes as well. Therefore, the following discussion is not limited to studies using NHST.

2.1. The role of sample sizes for power

There are several reasons why sample sizes matter. The first reason is that power depends on sample size. That means, the larger a sample, the higher is the probability to find an effect, given it really exists (Cohen, 1962; 1992). This observation is not new, and researchers have been aware of it for a long time (e.g., Cohen, 1962). Nevertheless, psychological studies still often suffer from low power (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Fraley & Vazire, 2014; Sedlmeier & Gigerenzer, 1989). For example, Bakker and colleagues estimate the typical power in psychological research to be around .35 for studies that compare two independent samples when assuming an effect size of $d = .50$ (2012). Button et al. estimate the median power in neuroscience to be around .21 (2013).

Whereas most researchers are aware that smaller samples have a higher likelihood of producing a miss (i.e. of not yielding a significant test although the effect exists), many are not aware of the seemingly paradoxical fact that smaller samples also have a higher likelihood of producing a false-positive (i.e. of yielding a significant test although the effect does not exist). Low power "… negatively affects the likelihood that a nominally statistically significant finding actually reflects a true effect" (Button et al., 2013). This is due to the fact that the positive predictive value PPV (the post-study probability that a claimed effect is true) depends on the