"Mirror, mirror, on the wall is there any evidence at all?" - Critical reflections on evidence in sport psychology research

Andreas Ivarsson

Halmstad högskola

Andreas Stenling

Umeå universitet & Halmstad högskola

Erik Lundkvist

Gymnastik & Idrottshögskolan

Two common objectives in sport and exercise psychology research are to determinate if there is a relationship between two or more variables or if an intervention is effective or not (i.e., casual effects). Based on results obtained from a research study we are often eager to conclude that there are strong (or weak) evidence for the proposed relationship or intervention. This procedure might seem straightforward but there are several problems and critical issues that influence researcher's assessments of the level of evidence. Unfortunately many researchers in the sport and exercise psychology field does not acknowledge these problems and critical issues when interpreting study results, which leads to flawed conclusions about the level of evidence (Ivarsson & Andersen, 2016). In this article we will: (a) highlight what we believe are some of the most critical issues in the sport and exercise psychology field for assessing the level of evidence, and (b) provide suggestions for how to deal with these issues.

The constant search for "Average Joe"

In most psychological studies population-level statistics are used to analyze the data. In this type of statistical procedure the results are based on mean values. One problem with performing analyses based on mean values is that the mean value does not represent any individual in the sample. Instead, we create an imagery person (we can call him "average Joe") that we base all our conclusions on. Because we always are interested in the actual participants scores the mean value will not tell us what we really want to know. To illustrate this potential shortcoming we use a hypothetical scenario. We are interested in investigating the level of stress symptoms in a small group of people (we can call them Malin, Sara, Andreas, and Henrik). We distribute a list of symptoms where the total score can be 0 (no

symptoms at all) to 12 (large amount of symptoms). The following scores were obtained: Malin = 2; Sara = 5; Andreas = 9; Henrik = 11. If we calculate the mean score for this small sample the value will be 6.75. If we should draw a conclusion based on this score it would probably be something like "the participants had moderate levels of stress symptom." But is this a fair conclusion? None of the participants are very close to this number so a better conclusion would probably be "there were very large differences in stress symptoms between the participants".

The haut for the little *p*

In many published articles a *p*-value smaller than 0.05 is automatically considered as evidence for a meaningful relationship between variables or differences between groups (Ivarsson, Andersen, Stenling, Johnson, & Lindwall, 2015). This is a limitation because the *p*-value will not tell us what we as researches often are interested in, namely the magnitude of the relationship or difference between two groups/time points (Andersen, McCullagh, & Wilson, 2007). Also, because the *p*-value is dependent of data that never is observed (Wagenmakers, 2007), it is problematic to base conclusions about evidence on this metric when it is the observed data we are interested in. To conclude "by itself the p-value does not provide a good measure of evidence regarding a model or hypothesis" (Wasserstein & Lazar, 2016, p. 132), and therefore other procedures should be used to evaluate the level of evidence in statistical testing (Ivarsson et al., 2015).

Time is a weapon of time

Most theories in sport and exercise psychology relate to some kind of process and thereby explicitly or implicitly stipulate that time is needed for that process to occur. A proper temporal design is crucial to accurately capture these processes, which makes it somewhat surprising that a vast majority of the published studies in sport and exercise psychology are cross-sectional. In addition, it is common to see researchers draw conclusions about these temporal processes based solely on cross-sectional findings (Hagger & Chatzisarantis, 2009). Although the amount of longitudinal studies have increased over the last couple of years; why researchers choose a specific time interval oftentimes seem to be an arbitrary decision rather than a reflection of the a priori hypothesized temporal processes. To illustrate this disconnection between theory and practice we use burnout as an example. Burnout is a popular outcome of mental ill-health in sport psychology research. The main symptom of burnout is severe exhaustion, which is a consequence of a prolonged period of stress (Shirom, 2005). However, this process is seldom examined; instead differences between two time points in self-reported burnout scores are used as a proxy measure of the burnout

process (Gustafsson, Lundkvist, Podlog, & Lundqvist, 2016). However, it is highly unlikely that differences in burnout scores between two time points can be used as a proxy for the burnout process (i.e., that prolonged stress leads to burnout) and any conclusion about the burnout process based on data from only one or two time points can be severely misleading. This example is merely one of many that illustrates a mismatch between theory and research practice. It is crucial that researchers in sport and exercise incorporate multiple aspects of temporality into the what, how, and why of their theories and design studies accordingly.

Arbitrary metrics and the real-world meaning of numbers

Although human behavior is an essential part of psychology there has been a rapid decline over the last decades of studies where behaviors are measured in psychological research. In a review of articles published in the Journal of Personality and Social Psychology from 1966 to 2006, direct observation of behavior decreased from 90% of studies in 1976 to less than 20% of studies in 2006 (Baumeister, Vohs, & Funder, 2007). Similar findings have been observed in sport and exercise psychology research where direct observation of behavior occurs in less than half of the published studies (Andersen et al., 2007). Instead of focusing on peoples behaviors, there is an overreliance in the field of sport and exercise psychology on introspective self-reports, hypothetical scenarios, and questionnaire ratings to capture the key variables of interest. The problems with relying on data generated from these sources are well-known, and one of the most severe is common method bias. Common method bias is a well-established phenomenon; one common observation is that associations between variables from the same source (e.g., self-report questionnaires) are stronger compared to associations based on data from different sources, and these associations are inflated due to common method bias (Podsakoff, MacKenzie, & Podsakoff, 2012). The sources of the bias can be related to various factors, such as the effects of response style, item wording, proximity and reversed items, item context, and person factors, such as mood and interpersonal context. Another problem with questionnaire ratings are related to the meaning of numbers obtained. Andersen et al. (2007) referred to this as a problem with arbitrary metrics, which relates to the fact that what the numbers actually mean and their implications are very difficult to interpret. Andersen et al. (2007) stated that:

> *"Sport and exercise psychology researchers use numbers extensively. Some of those numbers are directly related to overt real-world behaviors such as how high an athlete jumps or how far some object is thrown. Many of those measures, however, do not have such intimate connections to real-world performance or behavior. They often involve self-reports on inventories or*

> *surveys that are measuring (or attempting to measure) some psychological, underlying, or latent variables such as task and ego orientation or competitive state anxiety. What those scores on self-report inventories mean may be somewhat of a mystery if they are not related back to overt behaviors. (p. 664)."*

For example, if an athlete report 2.5 on a 5-point Likert scale intended to measure burnout, engagement, hope, motivation, passion, perceived coach behaviors, or some other commonly occurring construct in our field, what does that mean? If the primary outcome variable in a study is a self-report measure without any known link to overt behaviors, what have we actually learned from that study? How will such knowledge help us facilitate athletes or exercisers adherence, performance, or well-being? The short answer is; it won't. There is a pressing need to move beyond arbitrary metrics and measure abilities, achievements, and behaviors that have real-world meaning.

So where should we go from here?

In this article we have highlighted methodological issues that we argue needs more attention because they can negatively influence the accuracy of results in sport and exercise psychology research. Aggregated data analysis (i.e., the Average Joe phenomenon), overreliance on a single test statistic (i.e., the *p*-value), lack of attention to temporality in research designs, and the use of arbitrary metrics are all directly related to the conclusions researchers can draw from a study and they can have a major impact on the accuracy of the conclusions. To prevent these issues and to increase the likelihood of producing "best available evidence" researchers are recommended to:

1. Avoid a sole focus on between-person analyses (i.e., a focus on Average Joe).Instead, researchers should be explicit about whether effects are expected at the between-person level or within-person level and design studies to reflect that. Whether results obtained from between-person analyses aligns with results from within-person analyses is an empirical question and not something that merely should be assumed (cf. Simpson's paradox).
2. Use effect sizes or other magnitude measures (e.g., Bayes Factor) when interpreting the results from statistical analyses. Given that the *p*-value does not indicate the magnitude of an association, difference, or effect researchers need to use other indicators of magnitude. Established effect sizes within the context of interest is one approach to assess the magnitude of an association, difference, or effect.

3. State hypotheses that reflect the temporal process of interest and design studies based on the expected temporal process. Careful consideration of time can illuminate when and how processes unfold and provide valuable information to coaches about how they can help athletes improve their performance or to exercise instructors about how to work with sedentary individuals who wants to be more physically active.
4. Design studies that measure abilities, achievements, and behaviors that have real-world meaning (instead of just focusing on self-reported measures). Using metrics that are meaningful in themselves, such as objective performance measures in sports, number of injuries, time spent in physical activity or sleep quality as a proxy for stress (e.g., assessed with an accelerometer) allows for a straightforward translation of research findings into practical applications.

References

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport and Exercise Psychology, 29*, 664-672.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396–403.

Gustafsson, H., Lundkvist, E., Podlog, L., & Lundqvist, C. (2016). Conceptual Confusion and Potential Advances in Athlete Burnout Research. *Perceptual and Motor Skills. 123(3) 784–791.*

Hagger, M. S., & Chatzisarantis, N. L. D. (2009). Assumptions in research in sport and exercise psychology. *Psychology of Sport and Exercise*, *10*(5), 511–519.

Ivarsson, A., & Andersen, M. B. (2016). What counts as ”Evidence” in Evidence-Based practice? Searching for some fire behind all the smoke. *Journal of Sport Psychology in Action, 7*, 11-22.

Ivarsson, A., Andersen, M. B., Stenling, A., Johnson, U., & Lindwall, M. (2015). Things We Still Haven’t Learned (So Far). *Journal of Sport and Exercise Psychology, 37,* 449-461.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539–569.

Shirom, A. (2005). Reflections on the study of burnout. *Work & Stress*, *19*(3), 263–270. https://doi.org/10.1080/02678370500376649

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.