International Journal of Sports Physiology and Performance, 2021, 16, 1559-1560
https://doi.org/10.1123/ijspp.2021-0467

Human Kinetics
EDITORIAL

Coping With the "Small Sample–Small Relevant Effects" Dilemma in Elite Sport Research

Research in elite sport faces a characteristic tension between the inherently small number of elite athletes (and therefore lack of study participants) and the high relevance of even tiny differences. This characteristic constellation regularly leads to a mismatch of required and achievable sample size.

A suboptimal sample size not only increases the risk of failing to detect an effect of relevant magnitude (low statistical power) but also decreases the chance of a positive result to be a replicable, true discovery.[1] Furthermore, even for true positives, studies with small samples tend to overestimate the magnitude of the effect.[1,2] This "winner's curse" can be explained by the combination of 2 factors: On the one hand observed effect sizes are not stable with few data points. On the other hand small studies only provide significant results for the effect sizes that happen to be largest.[1,2]

Currently, the prevailing approach is to justify a suboptimal sample size by the elite status of the participating athletes. Alternatively, athletes with a lower performance level are included. However, the former is beset with the mentioned limitations and the latter questions the transferability to the elite level. Fortunately, a "standard" study plan (eg, a randomized controlled trial [RCT] analyzed by ANOVA or ANCOVA) can be adjusted on all levels—from project strategy over study design and methodology to data analysis and communication of results. Specific adjustments allow reducing the number of required participants (eg, by limiting random variation or gathering several observations per participant), as well as limiting the negative consequences of an eventually remaining mismatch (eg, by resampling and shrinkage techniques or Bayesian methods).

One priority is optimizing the signal-to-noise ratio, which is (in the form of a standardized effect size) an important determinant of the required sample size. Therefore, emphasis needs to be put on the reliability of measurement tools, standardization of protocols, and/or the optimal timing of tests.[1] Of note, the increase in the standardized effect size achieved by the intentional reduction of random variation in a study will not apply under the "normal" conditions of practical application. Therefore, fixed thresholds for standardized effect sizes (eg, a Cohen *d* of 0.2) may not always be helpful if sample size is truly to be based on the practical relevance of an effect. Rather, practical relevance should be judged based on relevant differences in practice (eg, competitions). In the context of sample-size calculation, the expected random variation in the measure can then be included, for example, based on reliability trials of previous work.

Another option for improvement consists in adjusting the study design to increase the number of data points while keeping the principles underpinning the success of the RCT (eg, control, blinding, randomisation). The key tweak here is repetition on the individual level. Depending on the reversibility of the effect, as well as on time, resource, and acceptability constraints, this approach may range from averaging replicated tests in an otherwise "normal" study, crossover trials, to replicated crossover and multiple baseline designs (aggregated single-subject designs). Note that such study designs lead to hierarchical data sets and have to be analyzed by appropriate techniques such as mixed effect models.

An important yet rarely explicitly discussed aspect is the strategic aim of the project: Is the focus on generalizable inference or, rather, on improved decision making in the sport? While the former is the habitual perspective of research scientists, the latter is arguably more prevalent when working with elite athletes. In this case what is called for is not the evidence the current data set can provide but a summary of the currently available evidence specific for the framework conditions in question. Therefore, formally including previous knowledge ("Bayesian updating") can offer a legitimate head start to a small data set.[1] Previous trials on a lower performance level or practitioner's experiential knowledge are 2 potential sources for valuable "prior" knowledge. Importantly, when combined with single-subject designs, valid individualized estimates (with appropriate precision) can be provided to the participating athletes.

Building on these foundations, a number of statistical techniques (eg, resampling and shrinkage techniques) can further improve insights from suboptimal sample sizes. While these approaches can offer decisive advantages in specific contexts, their selection and proper implementation require collaboration with an expert statistician. Note that highly multivariate analyses from the realm of machine learning and artificial intelligence rely on massive numbers of observations, which can rarely be realized in elite sports. These applications can be associated with a particularly high risk of unreplicable results ("overfitting," "scarce data bias"[3]). Importantly, versatility and adaptability are mirrored by the need for utmost transparency, including a predetermined study plan to avoid increasing the risk of false positive results due to (conscious or unconscious) "p-hacking." Finally, the restrictions of available samples make the replication of research results very important to combat the underpinning limitations.[4,5]

Sabrina Skorski, IJSPP *Associate Editor,*
Saarland University, Saarbrücken, Germany

Anne Hecksteden,
Saarland University, Saarbrücken, Germany

References

1. Hecksteden A, Kellner R, Donath L. Dealing with small samples in football research [published online ahead of print September 14, 2021]. *Sci Med Football*. doi:10.1080/24733938.2021.1978106
2. Senn S. Transposed conditionals, shrinkage, and direct and indirect unbiasedness. *Epidemiology*. 2008;19(5):1–2. PubMed ID: 18703929 doi:10.1097/EDE.0b013e318181b3e3

3. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *Br Med J*. 2016;352:i1981. PubMed ID: 27121591 doi:10.1136/bmj.i1981
4. Ioannidis JPA. Why replication has more scientific value than original discovery. *Behav Brain Sci*. 2018;41:e137. PubMed ID: 31064545 doi:10.1017/S0140525X18000729
5. Zwaan RA, Etz A, Lucas RE, Donnellan MB. Making replication mainstream. *Behav Brain Sci*. 2017;41:e120. PubMed ID: 29065933 doi:10.1017/S0140525X17001972