SPORTSCIENCE sportsci.org
Perspectives / Research Resources

# What's Behind the Numbers?
# Important Decisions in Judging Practical Significance

Greg Atkinson

Sportscience 11, 12-15, 2007 (sportsci.org/2007/ga.htm)
Research Institute for Sport and Exercise Sciences, Liverpool John Moores University, Liverpool L3 2ET, UK. Email.
Reviewer: Weimo Zhu, Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801.

In an applied field like sport and exercise science, inferences based on estimation of true effect sizes are usually more important than inferences about statistical significance. Inferences about estimation are conventionally made using confidence intervals, which are associated with several critical judgments. The most important decision concerns the smallest effect size that is practically or clinically important. A recently published new approach to sample size estimation also raises issues of judging the appropriate coverage probability of a confidence interval (e.g. 90 or 95%) as well as the degree of overlap between confidence limits and the smallest worthwhile effect. It is these *a priori* rationalized decisions that underpin the mathematics of confidence intervals, the probabilistic inferences made from them and associated issues like sample size estimation and claims that a statistical approach is too conservative or liberal. First, I discuss that the "null" in the null hypothesis testing process does not always need to be set at zero. If the smallest worthwhile effect itself is selected as the null value, then this process not so isolated from practical significance. Second, I contrast ideas on boundaries of overlap between confidence limits and the smallest worthwhile effect with other published guidelines on using confidence intervals to interpret study results. It is these differences in delimited probability coverage that govern the apparently lower sample sizes required for the new approach. Third, I illustrate how critical the decision on smallest worthwhile effect size can be for accuracy of study conclusions, and question whether uncertainty in this decision process might, in some instances, compromise the accuracy of the inferential statements that are made following statistical analysis. KEYWORDS: confidence intervals, null hypothesis, Type I and II statistical errors, smallest worthwhile effect.

Reprint pdf · Reprint doc

As he mentioned in his recent article (Hopkins, 2006), Will Hopkins' latest ideas about sample size estimation have arisen from a long-standing interest in the confidence interval approach to interpretation of study conclusions. Indeed, Will has been instrumental over the last two decades in communicating the advantages of such an approach amongst sport and exercise scientists. It is undeniable that confidence intervals help researchers to appraise the "real-world" relevancy of their study outcomes and that Will's spreadsheets are useful tools to help researchers make such an appraisal.

My personal interest in Will's article centers on the underpinning philosophy of the ideas rather than the mathematical accuracy of the spreadsheets derived from the "statistical first principles" which Will adopts. I know Will to be a highly competent mathematician who has a gift for communicating complicated mathematical concepts in a "researcher-friendly" way, especially through the use of his spreadsheets.

I think Will's claims that his new approach leads to sample sizes one third the size of "traditional methods" need to be viewed from a philosophical standpoint in order to unravel how this difference in numbers comes about. Such claims are especially interesting given that there are surprisingly tight relationships, both philosophically and mathematically, between some interpretations of the confidence interval approach and the null hypothesis testing process. For example, if the lower bound of a 95% confidence interval is exactly zero, then the exact P-value for statistical significance of the sample mean is 0.05 (5%). This makes sense,

since both the lower bound of the confidence interval and the P=0.05 in the null hypothesis testing process basically suggest that it is unlikely that the true population effect size is zero (or, put another way, that the observed effect size is unlikely to be merely due to chance sampling error). I know that Will is not too comfortable with this relationship between 95% confidence intervals and statistical significance in the null hypothesis testing process and I believe this is one reason why 90% confidence intervals are preferred by him and other statisticians.

I would like to make some comments, which may be relevant, about the "null" in the null hypothesis testing process. Firstly, the null value does not have to be set at zero. The null assumption can also be that the effect size is equal to the smallest worthwhile magnitude. "Null" in this sense means "not important" and suggests that the null hypothesis testing process is not completely disconnected from issues surrounding practical significance. I think adoption of this philosophy in the past would have at least reduced the instances of researchers automatically assuming that statistical significance is synonymous with practical importance. It is also not very well known that, as part of the philosophy of a one-tailed, directional analysis, the null hypothesis should be stated that the observed effect is zero or opposite in direction to that hypothesized by the researcher. This is because both these scenarios should result in the same study conclusion; the intervention should not be adopted.

Given Will's claims, it may surprise some readers when I say that there are some published interpretations of confidence intervals (e.g., Guyatt et al., 1995) which lead to estimations of *larger* (not smaller) sample sizes than for the null hypothesis testing procedure (when zero is the chosen null value). This is because the lower bound of a confidence interval might be larger than zero (hence the sample mean is statistically significant) but might not be larger than the smallest worthwhile effect. Some statisticians interpret this situation as the sample size not being large enough to be reasonably certain that the true population effect is larger than the smallest worthwhile effect, i.e. more subjects are needed to narrow the confidence interval and therefore arrive at a more precise conclusion. One can tell from the work Will has done on boundaries of benefit/harm that he is one of the statisticians that does not agree with this rather conservative pass-fail approach to confidence interval interpretation. Still, it serves to illustrate that the interpretation of confidence intervals is itself under debate, even without bringing in the Bayesians!

So, in view of the drastic reduction in estimated sample size, what exactly is Will doing differently in terms of the philosophy of applying probabilistic statements to study conclusions? If multiple assumptions have been made, how have these been rationalized? The answer to this latter question is especially important given the oft-cited criticism that the popular P<0.05 (5%) cut-off value for statistical significance in the null hypothesis testing process is quite arbitrary, although to be wrong about a claim of significance, given the observed data, only one time out of 20 seems a decent delimitation of "reasonably certain" to me.

Will believes that the use of the P<0.05 cut-off value is not only arbitrary but it leads to decisions that are too conservative. Is Will fighting a generalization with another (or several other) generalization(s) in this respect? Who or what is P<0.05 too conservative for? Doesn't such a view actually detract from what is really important - that the level of alpha (or indeed any delimitation about probability coverage or levels in data analysis) is a situation-specific delimitation? The P<0.05 cut-off could be viewed as too *liberal* in some circumstances, e.g. the use of an antiviral drug to combat HIV infection when that drug might have serious side effects. Will's solution to this problem seems to involve the introduction of two new types of decision error with delimited acceptable cut-off values of 0.5% and 25% (to be fair, Will cites these as examples). What is the exact rationale for these values? Following these delimitations, then the acceptable cut-offs for qualitative conclusions of "beneficial", "trivial", etc, are introduced. What should these probabilistic values be and what philosophical basis drives them? If Will's new methods are adopted, then all these situation-specific delimitations should come to the forefront of the researchers mind. Do we need discussion-based position statements to be formulated for all these delimitations which affect the study conclusion process?

Inherent in the confidence interval approach

to interpreting study conclusions is the most important delimitation a researcher needs to make; the selection of the smallest outcome magnitude that is clinically or practically important. Will maintains that any researcher who cannot arrive at such a value should "quit the field"! I can see his point in terms of the number of researchers who seem unable to even discuss the practical importance of their findings and agree that this inability is a terrible side effect of over-reliance on the null hypothesis testing approach. Nevertheless, I am not so sure that sport and exercise scientists have such an easy job in arriving at this smallest worthwhile effect.

Will maintains that a change of approximately 0.5 of the within-subject variability in performance between competitions is probably worthwhile for sports performance contexts (Hopkins et al., 1999). This cut-off value was arrived at following a study (the first of its kind) on the within- and between-athlete variability of real track-and-field performances at the elite level. Using these data, Will was able to estimate how much the within-athlete performance needs to change in order for it to make a difference in terms of winning places. But how does such a cut-off value relate to other scenarios, especially when such values have been calculated with all the variability associated with real-world situations? I am not challenging the delimitation here but wonder if we need to formalize the process of arriving at these decisions? Also, can such cut-off values derived from the real world be applied to the more tightly controlled environment of a laboratory experiment? For example, I have found recently that within-player variation (CV) of real soccer motion analyses can be as high as 100%. This variability is not surprising given the myriad of tactical and behavior variations between soccer matches. I don't think this magnitude of variability will be present if one researches an externally-valid component of soccer performance in the controlled environment of the laboratory. Will's value for a meaningful effect size of 0.5 x within-subject variability is at least better, in terms of underlying rationale, than Cohen's 0.2 of a between-subjects SD. How has this latter cut-off value been rationalized in terms of sports performance, physiology of exercise or indeed any outcome relevant to exercise science? Cohen was not a sport and exercise scientist, so he wasn't even in the field for him to be able to quit it!

Of course, the size of worthwhile effect should be an informed decision based on knowledge about what really makes a difference. But how easy is such a decision, especially when the study outcome variable is part of an overall concept? For example, what is the smallest difference in bowling speed that makes a difference to overall cricket performance of the team? This question was exactly the one Will needed to answer when he co-authored a recent paper (Petersen et al., 2004). In response to a training intervention, the smallest worthwhile change in bowling speed was stated by Peterson et al. to be 5 km/h as "the smallest that a top batsman would notice". Nevertheless, a smallest worthwhile effect size of 2.5 km/h was also stated as being "beneficial to a world-class bowler". As an illustration of how vital these decisions about smallest worthwhile effect are, and how clearly rationalized they should be, it was interesting that Peterson et al. found that the 90% confidence interval for the change in bowling speed was 1.2 to 4.2 km/h. This confidence interval tells us that a zero (null) change in true bowling speed is very unlikely (since the lower limit of the interval is 1.2). Nevertheless, the true change in bowling speed could be beneficial according to one delimited worthwhile effect (2.5 km/h) but not another (5 km/h), since the upper limit was higher than the former but lower than the latter delimited cut-off. Therefore, whilst Peterson et al. were pretty sure that the intervention induced an improvement in bowling speed, their study conclusion was less certain, according to their delimited worthwhile effect sizes. My question is to what extent should this ambiguity in the magnitude of the smallest worthwhile effect be built into Will's probabilities of "very likely beneficial", "trivial", etc? If the anchor between the delimited smallest worthwhile effect size and real world relevancy is pretty loose, is it actually worth being so precise with all the probabilities associated with the observed effect?

In summary, I believe that the most important issues in Will's article are not sample size calculations, but the new philosophy underpinning his new approaches to arriving at study conclusions using confidence intervals. There are new delimited conclusion error types and new boundaries of overlap between confidence

interval and smallest worthwhile effect. Will has set a very important ball rolling but its path needs to be clearly steered and agreed on in my opinion.

Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S (1995). Interpreting study results: confidence intervals. Canadian Medical Association Journal 152, 169-173

Hopkins WG, Hawley JA, Burke LM (1999). Design and analysis of research on sport performance enhancement. Medicine and Science in Sports and Exercise 31, 472-485

Hopkins WG (2006). Estimating sample size for magnitude-based inferences. Sportscience 10, 63-67

Petersen CJ, Wilson BD, Hopkins WG (2004). Effects of modified-implement training on fast bowling in cricket. Journal of Sports Sciences 22, 1035-1039

Published Dec 2007