SPORTSCIENCE · sportsci.org Latest issue
Perspectives / Research Resources This issue

Confidence Intervals and Meta-analysis Resolve the Replication Crisis

Will G Hopkins

Sportscience 27, 26-27, 2024 (sportsci.org/2024/ReplicationCrisis.htm)
Internet Society for Sport Science, Auckland, New Zealand. Email. Reviewer: Hans-Peter Wiesinger, Paracelsus Medical University (Institute of General Practice, Family Medicine and Preventive Medicine, Institute of Nursing Science and Practice; Salzburg, Austria.

Different studies sometimes reach different conclusions about the magnitude of the same effect, a phenomenon dubbed the replication crisis. Significance testing and sampling variation provide simple explanations for much of the apparent crisis, whereas compatibility or Bayesian interpretations of confidence intervals identify real replication failures. These failures are quantified as heterogeneity in random-effect meta-analysis, which can apportion at least part of the heterogeneity to the modifying effects of subject characteristics and study methodologies. Meta-analysis can also identify and discount heterogeneity arising from publication bias and sometimes from scientific fraud. Any remaining unexplained heterogeneity does not constitute a replication crisis.

Keywords: compatibility interval; confidence interval; magnitude-based inference; nil-hypothesis significance test; p value; precision of estimation.

Reprint pdf · Reprint docx · Slideshow

In November 2023 I visited my protégés in various European universities and presented a research seminar on the so-called replication crisis. The slideshow accompanying this article is a version of that seminar. What follows here is a summary. I recommend you view the slideshow as a full presentation to get the benefit of the extensive animations.

The notion of a replication crisis came to prominence following John Ioannidis' assertion that more than half the published claims of a "true relationship" based on statistical significance are false, because there is actually "no relationship" (Ioannidis, 2005). Goodman and Greenland (2007) subsequently criticized the claim as unfounded, but they agreed that "there are more false claims than many would suspect." A Wikipedia article provides an overview of the replication crisis and the reforms it has sparked. I think I deal here more succinctly with the roles of sampling uncertainty and meta-analysis in the resolution of the crisis.

It occurred to me that the crisis is largely illusory, a consequence of the usual misinterpretations of statistical significance and non-significance. Indeed, the crisis would largely resolve if researchers allowed for uncertainty in effects by interpreting outcomes using confidence intervals rather than the nil-hypothesis significance test. See my recent article for more on sampling uncertainty (Hopkins, 2022). When the uncertainties are such that two or more effects cannot be identical, there is replication *failure* rather than a replication *crisis*. Random-effect meta-analysis then deals with such failures by quantifying real difference in effects between studies as heterogeneity, which can be explained at least partly in a meta-regression by the modifying effects of study and subject characteristics that differ between study settings. For more on meta-analysis, see this article (Hopkins, 2018) and an earlier but recently updated article/slideshow (Hopkins, 2004).

Some of the real differences in effect magnitude between studies revealed by heterogeneity could be due to the insidious effect of statistical significance on the publication process, whereby studies of an effect are more likely to end up in journals if the effect is statistically significant. When a true effect is trivial or small, it will reach statistical significance in a study with a small sample size only if sampling uncertainty makes the observed effect larger than the true effect. Hence larger observed effects are more likely to reach statistical significance and therefore get published when sample sizes are small, whereas effects will be smaller in larger studies and may get published, even when they are non-significant. Minor manipulation of data to get a p value below the threshold for statistical significance ($p<0.05$) may also contribute to publication bias.

Adjusting for publication bias is therefore an important part of a meta-analysis, and I have updated the original slideshow with a description of two recent methods that in my own simulations work reasonably well in this respect. (For a review of all the methods, see Carter et al., 2019.) The meta-analyst should also exclude any studies where there is extensive fabrication of data, the evidence for which is sometimes revealed by unrealist errors of measurement.

In conclusion, John Ioannidis' claim of high prevalence of false relationships in the literature was misplaced but has led to valuable reforms of research methods. Replication failure rather than a replication crisis is inevitable in research and is readily explained by confidence intervals and heterogeneity in meta-analysis.

References

Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. Advances in Methods and Practices in Psychological Science 2, 115-144.

Goodman S, Greenland S. (2007). Why most published research findings are false: problems in the analysis. PLoS Medicine 4, e168.

Hopkins WG. (2004). An introduction to meta-analysis. Sportscience 8, 20-24.

Hopkins WG. (2018). Improving meta-analyses in sport and exercise science. Sportscience 22, 11-17.

Hopkins WG. (2022). Replacing statistical significance and non-significance with better approaches to sampling uncertainty. Frontiers in Physiology 13:962132, doi: 10.3389/fphys.2022.962132.

Ioannidis JP. (2005). Why most published research findings are false. PLoS Medicine 2, e124.

Published April 2024