

RESPONSE TO THE LETTER TO THE EDITOR

Replication: Do not trust your P -value, be it small or large

Colleagues have joined me in this Reply – we are grateful for the chance to consider the letter by Raiteri (2021) which covered one important part of Gandevia's editorial (Gandevia, 2021). The piece included one figure showing two curves in an attempt to deal with the interpretation of P -values and the replicability of an experiment based on P -values.

The intended messages from this component of the editorial were the following. First, the likelihood to obtain $P < 0.05$ in a replication increases markedly if very low P -values are obtained in the *initial* experiment (say below 0.001) (see Fig. 1 in Gandevia, 2021). Second, any initial P -value around 0.05 does *not* signify any trend at all because the probability of an exact replication to obtain $P < 0.05$ is approximately 50% (see below). This message is intended to deflate any excitement accompanying

an initial P -value just below 0.05 as an indication of significance. Perhaps you could equally think of such a value as a trend for *in-significance*. This point is not contested. Interestingly, for an initial P -value around 0.01 the probability of obtaining $P < 0.01$ in a replication is also approximately 50%. Third, the prediction intervals for a replication are extraordinarily wide – almost unbelievably so. For an initial experiment which obtains $P = 0.05$, the 80% interval for replication P -values is huge: from 0.0002 to 0.65 (for a two-tailed test; see Fig. 1). (It was given for a one-tailed test in the original graph with a lower, but still huge, interval from 0.00008 to 0.44 (Gandevia, 2021)). The above values for the prediction intervals are derived from Cumming (2008; Appendix B). They assume exact replications that are identical with the initial experiment, except using a new sample. It is not required to assume, as Raiteri (2021) stated, that the initial experiment estimated the true effect size exactly. In summary, the prediction inter-

vals for replication P -values, the grey bars in Fig. 1, arise from idealized exact replications but, even so, are alarmingly wide.

Now to some clarification of the first and third messages.

For the first message, the data in the original figure (Gandevia, 2021) were taken from Goodman (1992) for $P = 0.001$ –0.10 and from Curran-Everett (2016) for $P < 0.001$. They relate to an attempted exact replication, notwithstanding the acknowledged difficulty in such an attempt. The original text neglected to say that the probabilities were for the preferred two-tailed replication. Goodman's data for $P = 0.001$ –0.10 were chosen deliberately from the right column of his Table 1 because it used 'the more realistic scenario that we do not know that μ [true effect size]... and our uncertainty is modelled as a prior probability of μ that is locally uniform, that is the distribution of μ prior to the second experiment is proportional to its likelihood function in the first experiment'. It deserves mention that the general shape

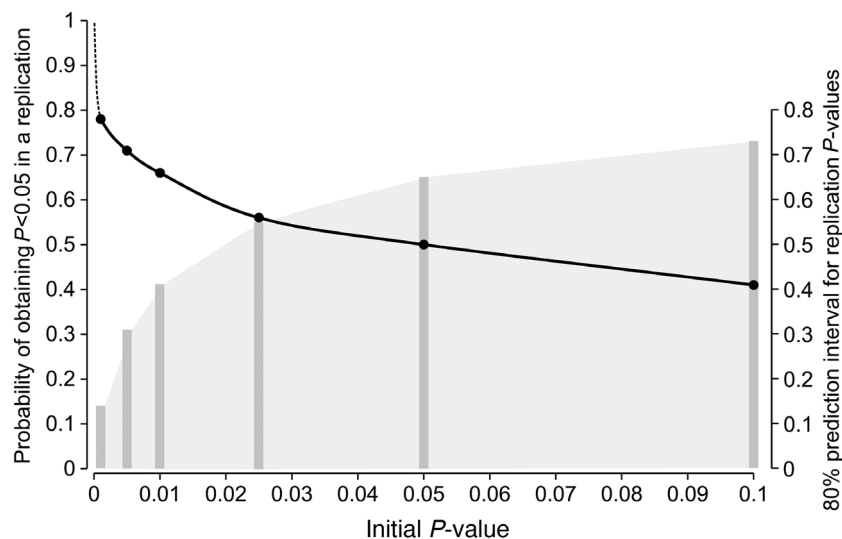


Figure 1. The neglected curves: probability that an exact, perfectly executed replication study will obtain $P < 0.05$, plus 80% prediction intervals for the P -value given by a replication

The continuous line in Fig. 1 shows the probability (see left axis for values) that a replication study will obtain $P < 0.05$ for a given P -value in the initial study (data from Goodman, 1992), with extrapolation dotted. For example, if an initial experiment obtains $P = 0.05$, there is only a 50% chance that an attempted exact replication will obtain a $P < 0.05$. Even if the initial $P = 0.01$, the chance is only 66% that an attempted replication will obtain $P < 0.05$ (i.e. a one-third chance to obtain $P > 0.05$). Grey bars represent the intervals that include the P -value given by a replication with an 80% chance (see right y-axis for values). The shaded area allows this prediction interval to be read off for initial P -values up to 0.1. Data are derived from the formula in Cumming (2008). For example, if an initial experiment generates $P = 0.05$, the 80% prediction interval for replication experiments will range from $P = 0.0002$ to 0.65 in idealized studies without imperfections. Realistically, this large range may even be an underestimate (see text). One of us (S.G.) still maintains that, for physiologists, teaching the messages from this pair of curves should be as important as teaching the haemoglobin dissociation curve

of the relationship between the P -value from the original experiment and the probability of its replication (at say 0.05) is largely independent of sample size and also the statistic used (Boos & Stefanski, 2011). Further, the approximate shape can be predicted by a number of statistical approaches (Boos & Stefanski, 2011; see also Cumming, 2008). A practical corollary is that when a replication is attempted, a large P -value can become small, and a small P -value can become large (e.g. Amrhein *et al.* 2019b)! This is illustrated formally in Fig. 1 in which the wide prediction intervals are clear even for studies in which the P -value from the initial experiment is very low (say <0.001). Further, these prediction intervals are not influenced by the sample or effect size, provided only that the sample is not too small (Cumming, 2008).




For the third message, arguments from Miller & Schwarz (2011) are cited by Raiteri (2021) and used to paint an even gloomier view about the difficulty of predicting replication P -values. They model the research 'world' as more complex than the simple one a physiologist may encounter in an attempted exact replication – not all parameters including factors related to the research field, and the individual experimenters, are known. Thus, we can expect that P -values will vary even more widely than in Fig. 1, and perhaps even sufficiently to justify Miller and Schwarz's conclusion that 'accurate estimates of replication probability are generally unattainable'.

Individual researchers will have to decide whether to consider the reproducibility of their results based on the massive range of replication intervals (Fig. 1) and the issues raised by Miller & Schwarz (2011). Alternatively, they can use different approaches: for example, the use of confidence intervals and meta-analysis from repeated experiments (Cumming & Calin-Jageman, 2017). They can reinterpret and re-express the observed P -values based on the concept of a false discovery rate (Colquhoun, 2014, 2017). Finally, they can take some reassurance in the (rare) event that should the initial study have a P -value below 0.05 and this probability is achieved in a replication, then the likelihood of a 'true' finding is substantially increased.

Our discussion indicates that, unless P -values are extremely small, they give very little or essentially no information (Miller & Schwarz, 2011; Raiteri, 2021) about what P -value a replication is likely to obtain. More broadly, a P -value, as

a single number, cannot make salient the degree of uncertainty in a result, for example because it mixes information on the size of the effect and how precisely it was measured (Amrhein *et al.* 2017). In contrast, the extent of a confidence interval (CI), which some prefer to label a compatibility interval (Amrhein *et al.* 2019a,b), makes that uncertainty salient: a short CI gives reassurance, a long CI gives a disappointing message that there is much uncertainty in knowledge of the effect size being investigated. On average, there is an 83% chance that a replication result (point estimate) falls within a 95% CI (Cumming, 2014). Thus, the observed 95% CI provides on average an 83% prediction interval for a replication result, although we should not forget the 'dance of the confidence intervals' (Cumming, 2014): this shows how a valid interval will change from sample to sample due to random variation (Amrhein *et al.* 2019a,b). But, in contrast to a P -value, a CI directly uncovers a range of values that replication estimates could take.

Scrutinizing results through replication plus using other techniques and arguments to reach conclusions is crucial to the progress of physiology, indeed all science. It should not be obstructed by the misunderstanding and misuse of P -values and other statistical measures.

Simon Gandevia¹ , Geoff Cumming² ,
Valentin Amrhein³ and Annie Butler⁴ 

¹Neuroscience Research Australia, Sydney, NSW, 2031, Australia

²School of Psychology and Public Health, La Trobe University, Melbourne, Victoria, 3086, Australia

³Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland

⁴Neuroscience Research Australia, Sydney, NSW, 2031, Australia

Email: s.gandevia@neura.edu.au

Edited by: Kim Barrett & Ian Forsythe

Linked articles: This is a reply to a Letter to the Editor by Raiteri. To read the Letter to the Editor, visit <https://doi.org/10.1113/JP281472>. These Letters refer to an Editorial by Gandevia. To read the article, visit <https://doi.org/10.1113/JP281360>.

References

- Amrhein V, Greenland S & McShane B (2019a). Retire statistical significance. *Nature* **567**, 305–307.

Amrhein V, Korner-Nievergelt F & Roth T (2017). The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ* **5**, e3544.

Amrhein V, Trafimow D & Greenland S (2019b). Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *Am Stat* **73**(sup1), 262–270.

Boos DD & Stefanski LA (2011). P -Value precision and reproducibility. *Am Stat* **65**, 213–221.

Colquhoun D (2014). An investigation of the false discovery rate and the misinterpretation of p -values. *R Soc Open Sci* **1**, 140216.

Colquhoun D (2017). The reproducibility of research and the misinterpretation of p -values. *R Soc Open Sci* **4**, 171085.

Cumming G (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* **3**, 286–300.

Cumming G (2014). The new statistics: why and how. *Psychol Sci* **25**, 7–29.

Cumming G & Calin-Jageman R (2017). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge, Abingdon, UK.

Curran-Everett D (2016). Explorations in statistics: statistical facets of reproducibility. *Adv Physiol Educ* **40**, 248–252.

Gandevia SC (2021). Publications, replication, and statistics in physiology plus two neglected curves. *J Physiol* **599**, 1719–1721.

Goodman SN (1992). A comment on replication, p -values and evidence. *Stat Med* **11**, 875–879.

Miller J & Schwarz W (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychol Methods* **16**, 337–360.

Raiteri BJ (2021). The unknowable probability of replication. *J Physiol*.

Additional information

Competing interests

No competing interests declared.

Author contributions

All authors have read and approved the final version of this manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All persons designated as authors qualify for authorship, and all those who qualify for authorship are listed.

Funding

None.

Keywords

p -values, replication, reproducibility, statistics