



Why Underpowered Studies Keep Passing Peer Review in Quantitative Biology

Emilio Harutyunyan^{1*}, Thabo Jimenez², Dragan Berg³

¹Department of Life Sciences, Siberian School of Life Sciences, Philippines

²Department of Life Sciences, Netravati Institute of Life Sciences, South Africa

³Department of Life Sciences, Sharavati Research Institute of Life Sciences, Jordan

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijbb.2026.v1.i1.pp27-32](https://doi.org/10.62715/ijbb.2026.v1.i1.pp27-32)

Received: 26 December 2025

Accepted: 21 February 2026

Published: 11 April 2026

Abstract

That many biological studies are too small to detect the effects they look for has been shown repeatedly over six decades, and the demonstrations have changed little. This Perspective asks why peer review, the checkpoint at which a design flaw of this kind might be expected to be caught, so rarely catches it. The argument is that low power is close to invisible in a finished manuscript. Reviewers see a result and not the planning behind it, and a statistically significant result is widely read as proof that the sample was large enough, when at low power it is better read as a warning that the estimate is exaggerated. Reviewers are also drawn from the community whose conventional sample sizes are in question, and in cell and imaging work the unit of replication is often left ambiguous, so that thousands of measurements obscure a handful of independent experiments. These conditions feed a loop in which inflated published effects become the basis for planning the next small study. The article considers the main objections, including ethical limits on animal numbers and the exploratory character of much laboratory work, and argues that the remedy lies in moving scrutiny of design to an earlier point and in asking for different information at submission.

Keywords: Statistical power, Peer review, Sample size justification, Effect size inflation, Reproducibility, Registered Reports

© 2026 Emilio Harutyunyan et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A study is underpowered when its sample is too small to give a reasonable chance of detecting an effect of the size that would matter if it existed. The consequences are well understood. True effects are missed more often than not, and the effects that do reach significance are overestimated, sometimes with the wrong sign (Gelman and Carlin, 2014). When most studies in a field have low power, a significant finding is weaker evidence than its p-value suggests, a point central to the argument of Ioannidis (2005) about the reliability of published research.

None of this is news to statisticians or to most working biologists. What needs explaining is the persistence of the practice in the face of that knowledge, and in particular the role of peer review. Every

* Corresponding author: Emilio Harutyunyan, Department of Life Sciences, Siberian School of Life Sciences, Philippines. E-mail: emilio.harutyunyan@siberian.edu

© 2026 Emilio Harutyunyan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

underpowered paper in the literature was read before publication by an editor and by reviewers chosen for their expertise, and it was accepted. This article argues that review fails here for structural reasons and not through carelessness, that the reasons are especially strong in quantitative biology, and that they point to remedies different from another exhortation to reviewers to check the sample size.

2. An Old Problem With a Stable Record

The first systematic audit of power was carried out by [Cohen \(1962\)](#), who worked through a volume of a psychology journal and found that the studies in it had roughly even odds of detecting an effect of medium size. A generation later [Sedlmeier and Gigerenzer \(1989\)](#) repeated the exercise on a later volume of its successor journal and found no improvement, and asked in their title whether studies of power have any effect on the power of studies. [Button et al. \(2013\)](#) brought the question to neuroscience, using published meta-analyses to estimate the power of the contributing studies, and concluded that it was typically far below the conventional target of 0.80.

Laboratory biology has its own evidence. A survey of published animal experiments found that the basis for the chosen sample size was hardly ever reported, alongside frequent omissions in the description of design and analysis ([Kilkenny et al., 2009](#)). When the reasoning behind a sample size is absent from the record, nobody downstream can evaluate it. The stability of these findings across decades and disciplines suggests that whatever maintains small samples is not addressed by pointing the problem out.

3. Why Review Does Not Stop It

Four features of the review process work against the detection of inadequate power. They are summarised in Table 1, together with two further mechanisms that operate before a manuscript ever reaches a reviewer.

3.1. The information is not on the page

Power is a property of a design considered before the data exist. It depends on the sample size, the variability of the measurement and the smallest effect the investigator cares about. A conventional manuscript reports the first, allows the second to be estimated, and is silent on the third. A reviewer who wants to judge power must supply an effect size of interest personally, which few feel entitled to do for someone else's question. The alternative that some authors and reviewers reach for, computing power from the observed effect, is circular: observed power is a restatement of the p-value and adds nothing to it ([Hoenig and Heisey, 2001](#)).

3.2. Significance looks like its own justification

The most common defence of a small study is that it found something. If the effect reached significance with so few samples, the reasoning goes, the effect must be large and the sample was evidently enough. The intuition is strong and it is mistaken. [Tversky and Kahneman \(1971\)](#) documented it among trained researchers, who expected small samples to mirror the population far more closely than they do. At low power the p-value itself varies widely from one replicate to the next ([Halsey et al., 2015](#)), and the estimates that happen to cross the threshold are those that chance has pushed upward. A reviewer faced with a significant result from a small sample is looking at the situation in which exaggeration is most likely, and is being invited to see it as the situation in which concern about sample size is least warranted.

3.3. Reviewers belong to the field

Peer review compares a manuscript with the standards of its field, and reviewers are qualified by having met those standards themselves. Where three animals per group or three independent cultures is the custom, a submission that follows the custom gives a reviewer nothing to object to, and objecting would implicitly condemn the reviewer's own publications. Review is therefore effective at enforcing a convention and ineffective at questioning one. A convention about sample size is also unlike a convention about controls, because its failure produces no visible anomaly in any single paper.

Table 1: Mechanisms that allow underpowered studies through peer review, and a countermeasure for each

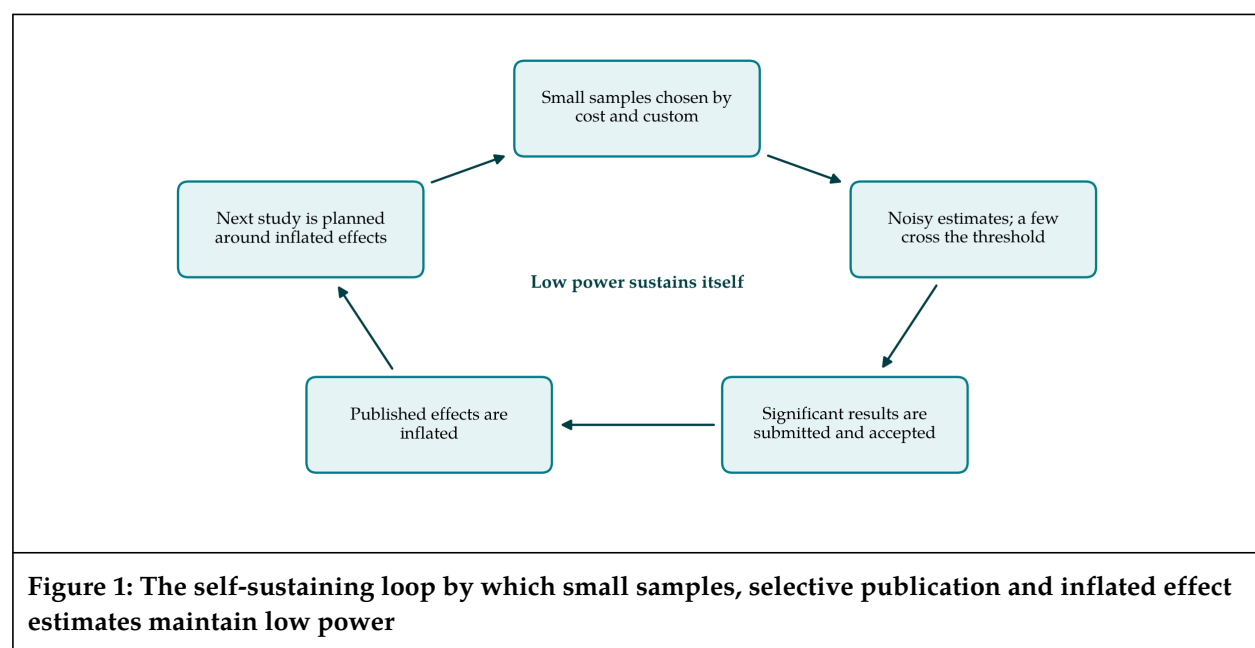
Mechanism	How it operates	Possible countermeasure
Missing information	The manuscript states the sample size but not the effect it was planned to detect	Require a sample size justification (Lakens, 2022)
Significance as vindication	A significant result is taken to show that the sample was adequate	Ask how exaggerated the estimate could be (Gelman and Carlin, 2014)
Shared norms	Reviewers use the same conventional sample sizes in their own work	Statistical review and reporting checklists (Percie du Sert et al., 2020)
Ambiguous unit	Cells or measurements are counted where independent experiments should be	Require the experimental unit to be stated (Lazic et al., 2018)
Selective submission	Small studies with null results are never sent for review	Acceptance before results are known (Chambers, 2013)
Incentives	Many small studies yield more publishable results than a few large ones	Reward design quality in evaluation (Smaldino and McElreath, 2016)

3.4. The unit of replication is easy to blur

Quantitative biology adds a difficulty of its own. Modern instruments return very large numbers of measurements: thousands of cells per image, hundreds of trajectories per cell, many time points per trace. A figure legend that reports the number of cells conveys an impression of abundant data, while the number of independent experiments, animals or cultures behind them may be two or three. [Lazic et al. \(2018\)](#) set out why only the independently treated unit counts towards the sample size for a treatment comparison and how often manuscripts leave that unit unclear. A reviewer cannot identify an underpowered design when the manuscript does not make plain what the sample size is.

4. A Loop That Sustains Itself

Two mechanisms act before review and determine what reviewers see. Small studies that find nothing tend not to be written up, the file drawer effect described by [Rosenthal \(1979\)](#), and flexibility in analysis raises the proportion of small studies that find something ([Simmons et al., 2011](#)). The submissions that arrive are thus a filtered set in which small samples and significant results go together.



The published effects are then used to plan new work. An investigator who bases a sample size on the effect reported in the literature, as textbooks advise, will calculate that a small sample is sufficient, because the reported effect is inflated. The loop in Figure 1 closes without anyone in it behaving unreasonably by local standards. Modelling work has formalised the wider dynamic. [Smaldino and McElreath \(2016\)](#) showed that when laboratories are rewarded for the number of positive findings, methods that generate such findings cheaply spread through a population of laboratories without any intent to deceive, and [Higginson and Munafò \(2016\)](#) found that the rational strategy for a scientist under such incentives is to run many small studies.

5. Objections

Several objections deserve an answer. The first is ethical. Animal experiments are bound by the principle of reduction, and demanding larger groups appears to conflict with it. The conflict is only apparent. An experiment too small to answer its question uses animals to no purpose, and the ARRIVE 2.0 guidelines accordingly ask authors to explain how the sample size was decided, not to minimise it ([Percie du Sert et al., 2020](#)).

The second objection is that much laboratory work is exploratory, and that requiring power calculations for exploration is a category mistake. This has force. Exploration needs no prior hypothesis about effect size. The difficulty is that exploratory studies are seldom reported as such: they are written up with confirmatory language and tested against confirmatory thresholds. A study presented as hypothesis generating, with estimates and intervals in place of claims of discovery, is not open to the criticism made here.

The third is that biological effects are often large, so small samples suffice. Sometimes they are, and a knockout that abolishes a signal needs few replicates to show it. But the claim has to be made and defended in advance for the effect in question, which is what a sample size justification is. A fourth objection holds that meta-analysis will eventually correct individual small studies. It cannot do so when the studies that enter it have been selected for significance.

6. What Could Change at the Point of Review

If the analysis above is right, asking reviewers to be more vigilant will achieve little, because the information they need is missing and the cues they have point the wrong way. Three changes address the mechanisms directly. The first is to move review of the design ahead of the results. In the Registered Reports format introduced by [Chambers \(2013\)](#), the rationale, methods and planned sample are reviewed and provisionally accepted before data are collected, so that the sample size is examined when it can still be changed and when no result exists to make it look adequate. The format is now offered by many journals and was among the measures recommended in the manifesto of [Munafò et al. \(2017\)](#).

The second change is to alter what a conventional submission must contain. A justification of sample size need not be a formal power analysis. [Lakens \(2022\)](#) describes several legitimate forms, including justification by resource constraints, provided that the consequences for what the study can show are stated. Coupled with a required statement of the experimental unit and the number of independent replicates, this puts the missing information on the page. The third change is to shift the question reviewers ask of a significant result from whether an effect exists to how well it has been estimated, which makes a wide interval from a small sample a visible limitation. These are modest demands, and each answers one of the mechanisms in Table 1.

Editors are better placed than reviewers to make such requirements stick. A reviewer who asks for a sample size justification in one manuscript is making a personal demand that the next reviewer may not repeat, and authors learn to treat it as the luck of the draw. A journal that asks for it in every submission, in a fixed place and a fixed form, turns it into a property of the venue. The same holds for the experimental unit: a mandatory line in the methods stating what was independently treated, and how many such units there were, costs authors a sentence and spares reviewers the task of reconstructing it from figure legends. None of these measures removes the incentives that favour small studies, and

journals acting alone cannot do that. They would, however, stop review from certifying designs it has had no means of examining.

7. Conclusion

Underpowered studies pass peer review because review, as ordinarily conducted, is not positioned to stop them. The reviewer arrives after the design is fixed, is shown a result that appears to vindicate it, shares the conventions that produced it, and in quantitative biology may not be told plainly how many independent observations there were. Sixty years of audits have shown that naming the problem does not shift it. Changing when the design is reviewed and what authors must state about it is more likely to, and both changes are within the power of journals to make.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. and Munafò, M.R. (2013). [Power failure: Why small sample size undermines the reliability of neuroscience](#). *Nature Reviews Neuroscience*, 14(5): 365-376. <https://doi.org/10.1038/nrn3475>
- Chambers, C.D. (2013). [Registered Reports: A new publishing initiative at Cortex](#). *Cortex*, 49(3): 609-610. <https://doi.org/10.1016/j.cortex.2012.12.016>
- Cohen, J. (1962). [The statistical power of abnormal-social psychological research: A review](#). *Journal of Abnormal and Social Psychology*, 65(3): 145-153. <https://doi.org/10.1037/h0045186>
- Gelman, A. and Carlin, J. (2014). [Beyond power calculations: Assessing Type S \(sign\) and Type M \(magnitude\) errors](#). *Perspectives on Psychological Science*, 9(6): 641-651. <https://doi.org/10.1177/1745691614551642>
- Halsey, L.G., Curran-Everett, D., Vowler, S.L. and Drummond, G.B. (2015). [The fickle P value generates irreproducible results](#). *Nature Methods*, 12(3): 179-185. <https://doi.org/10.1038/nmeth.3288>
- Higginson, A.D. and Munafò, M.R. (2016). [Current incentives for scientists lead to underpowered studies with erroneous conclusions](#). *PLoS Biology*, 14(11): e2000995. <https://doi.org/10.1371/journal.pbio.2000995>
- Hoenig, J.M. and Heisey, D.M. (2001). [The abuse of power: The pervasive fallacy of power calculations for data analysis](#). *The American Statistician*, 55(1): 19-24. <https://doi.org/10.1198/000313001300339897>
- Ioannidis, J.P.A. (2005). [Why most published research findings are false](#). *PLoS Medicine*, 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D., Hutton, J. and Altman, D.G. (2009). [Survey of the quality of experimental design, statistical analysis and reporting of research using animals](#). *PLoS ONE*, 4(11): e7824. <https://doi.org/10.1371/journal.pone.0007824>
- Lakens, D. (2022). [Sample size justification](#). *Collabra: Psychology*, 8(1): 33267. <https://doi.org/10.1525/collabra.33267>
- Lazic, S.E., Clarke-Williams, C.J. and Munafò, M.R. (2018). [What exactly is 'N' in cell culture and animal experiments?](#) *PLoS Biology*, 16(4): e2005282. <https://doi.org/10.1371/journal.pbio.2005282>
- Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J. and Ioannidis, J.P.A. (2017). [A manifesto for reproducible science](#). *Nature Human Behaviour*, 1: 0021. <https://doi.org/10.1038/s41562-016-0021>
- Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M.T., Baker, M. et al. (2020). [The ARRIVE](#)

- guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biology*, 18(7): e3000410. <https://doi.org/10.1371/journal.pbio.3000410>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3): 638-641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Sedlmeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2): 309-316. <https://doi.org/10.1037/0033-2909.105.2.309>
- Simmons, J.P., Nelson, L.D. and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11): 1359-1366. <https://doi.org/10.1177/0956797611417632>
- Smaldino, P.E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9): 160384. <https://doi.org/10.1098/rsos.160384>
- Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2): 105-110. <https://doi.org/10.1037/h0031322>