LINK TO ORIGINAL ARTICLE
LINK TO AUTHOR'S REPLY

Experimental power comes from powerful theories — the real problem in null hypothesis testing

John C. Ashton

In their excellent article, Button *et al.* (Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013))[1] go a long way towards identifying the solution to the crisis in neuroscience null hypothesis testing. However, although null hypothesis testing as originally conceived (and as is still used in applied research) is a powerful tool for decision-making, the problem is that in much of science, null hypothesis testing is no longer performed in the original manner.

In applied research, an effect size is specified in advance and defines a threshold for decision-making. Power analyses of a specific desired effect size may then be carried out before an experiment in which a null hypothesis is tested against an alternative hypothesis that is based on this desired effect size. As a result, hypotheses in applied research are highly testable. By contrast, in much of science, the alternative hypothesis is left open, with the only effect size under consideration being the one estimated from the data, so there is no standard upon which to compare this measured effect size, and therefore no hard basis for decision-making. In null hypothesis testing, any effect — no matter how small — may end up being statistically significant if enough replicates are used[2,3]. This means that the vague 'open' hypotheses of much of neuroscience are barely testable, as more replicates could always be added and more subtle effects searched for.

Although the advice to increase sample sizes and statistical power is sound, when combined with the notion that neuroscientists should search for ever more subtle effects, following this advice would mean that hypotheses in neuroscience become virtually untestable. If we fail to find a 10% effect, then we can always fall back on the possibility that with increased power, we might detect a 1% effect or 0.1% effect *ad infinitum*. Therefore, the advice may ultimately do little to restrain the proliferation of poorly testable (that is, hard-to-refute) hypotheses.

The only way to resolve this dilemma while retaining the advantages of traditional null hypothesis testing is to be specific about the theoretical predictions that our experiments are designed to test[4]. Whether the alternative hypothesis is designed to test for a subtle or strong effect depends entirely on the theory and problem under investigation. For some theories, only strong effects are relevant, but for others, subtle effects may be meaningful. If instead neuroscience becomes a needle-in-a-haystack search for ever more subtle effects irrespective of the presence of an explanatory theory, then neuroscience could degenerate into explanationless collections of observations — that is, mere 'stamp collecting'. The observations will be the fruit of virtually untestable hypotheses. These hypotheses will be retained if they are below a threshold for statistical significance — in the hope of detecting increasingly subtle effects — but sanctioned by a spurious statistical significance if this open-ended search for ever more subtle levels of effect lead to a type I error.

The solution to the problem is to increase discipline not only in analysis and experimental design but also in relating experiments to explanatory theory[5]. Much current practice instead seems to be an open-ended search for associations, reminiscent of old-style inductionism while superficially following the conventions of hypothetico-deductivism.

John C. Ashton is at the Department of Pharmacology & Toxicology, School of Medical Sciences, University of Otago, Dunedin 9140, New Zealand.

e-mail: john.ashton@otago.ac.nz

doi:10.1038/nrn3475-c2
Published online 3 July 2013

1. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013).
2. Carver, R. P. The case against statistical significance testing. *Harvard Educ. Rev.* **48**, 378–399 (1978).
3. Cohen, J. The earth is round ($p < .05$). *Am. Psychol.* **49**, 997–1003 (1994).
4. Meehl, P. E. Theory-testing in psychology and physics: a methodological paradox. *Philos. Sci.* **34**, 103–115 (1967).
5. Meehl, P. E. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* **46**, 806–834 (1978).

Competing interests statement
The author declares no competing financial interests.