LINK TO ORIGINAL ARTICLE
LINK TO CORRESPONDENCE BY QUINLAN
LINK TO CORRESPONDENCE BY ASHTON
LINK TO CORRESPONDENCE BY BACCHETTI

Confidence and precision increase with high statistical power

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò

We are delighted that our Analysis article (Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14**, 365–376 (2013))[1] has stimulated debate about the issues arising from low statistical power in neuroscience studies. Here, we take the opportunity to respond to some important points made by Quinlan (Misuse of power: in defence of small-scale science. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c1 (2013))[2], Ashton (Experimental power comes from powerful theories — the real problem in null hypothesis testing. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c2 (2013))[3] and Bacchetti (Small sample size is not the real problem. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c3 (2013))[4], and clarify our position on certain key issues raised in our original article[1].

Quinlan[2] draws a comparison between our observations and those made by Friston in his recent article[5]. However, we note that subsequent commentaries by Ingre[6] and by Lindquist and colleagues[7] echo many of the concerns that we raised and that Friston[8] agreed that he was "convinced by most of their observations". The main concern — originally raised by Friston[5] and reiterated by Quinlan[2] — is that high-powered studies will generate formally statistically significant differences for a 'trivial' effect. There are three important rejoinders to this concern. First, in studies with low statistical power, an observed effect will necessarily be large if it is to pass a pre-specified *p*-value threshold (typically 0.05), but this does not mean that the true effect will be large, or even exist at all. Second, the concern only applies if a *p*-value threshold (that is, typically 0.05) is used to either reject or (implicitly) accept the null hypothesis. As Ingre notes[6], larger studies actually protect against inferences from trivial effect sizes by allowing a better estimation of the magnitude of the true effect. A shift in emphasis away from significance testing towards the use of effect sizes and confidence intervals would therefore improve matters — this is a point we return to below. Third, the true effect size is not known in advance; what is considered 'trivial' can only be determined when the effect size is known. For example, candidate gene studies are frequently quite small *N* studies, with the implicit assumption that common genetic variants will be associated with reasonably large effects. Over a decade of relatively fruitless research showed that this assumption is not correct; the magnitude of these effects turns out to be much smaller[9], as has been borne out by more recent genome-wide association studies (GWASs), which are explicitly designed to be able to detect very small effects. Have these smaller effects turned out to be trivial? Not at all — in many cases, the genes that have been identified using well-powered GWAS methods have led to renewed interest in the role of the pathways implicated and have thereby generated genuinely new insights into disease mechanisms[10]. Moreover, learning that the effect sizes of common genetic variants on complex traits are very small has provided important fundamental insights into the genetic architecture of these traits[11]. Therefore, high power provides greater precision in the estimation of the actual effect size so that researchers can assess their importance or triviality with confidence.

Ashton[3] makes the important point that specification in advance of the effect size being sought is rare, which precludes the use of a priori power analysis. He argues that we should be more "specific about the theoretical predictions that our experiments are designed to test", which should include a prediction regarding the magnitude of the expected effect. This is related to our suggestion of placing greater emphasis on effect size and confidence intervals than on significance testing. Unfortunately, the use of significance testing in the absence of any mention of effect size, confidence intervals or prospective power remains the norm[12]. Some may argue that effect size is not relevant to the theoretical models they wish to test. That may be true if the models are imprecise about effect sizes. However, even in that case, data from low-powered studies are not useful for testing a theoretical model because they provide little opportunity to find conclusive evidence for or against a model and therefore provide limited scope for model refinement.

Bacchetti[4] argues that if one could take care of all the associated biases that have been empirically documented to be far more prevalent in very small studies than in larger studies, then small studies (or even very small studies) would be unproblematic. We agree that it would be wonderful if small studies and their research environment were devoid of biases and if all small studies on a particular question of interest could be perfectly integrated. However, this has not happened; to achieve this would require a major restructuring of the incentives for publishing papers, and especially for publishing novel and positive findings[13]. Although we applaud efforts to reduce biases, we believe that some larger, more definitive studies are also needed to address the problems we describe in our original article. For example, the 'winner's curse' will remain a problem for small studies even if all biases are eliminated. We sympathize with the view that studies may not need to be enormous and that studies with modest (but not very small) sample sizes may occasionally have some advantages over studies with large sample sizes[14]. Also, in some cases (for example, when a disease is rare or when sample size is unavoidably constrained), small studies may be the best we can achieve, and the optimal design might include choices for type I and type II errors that do not necessarily correspond to the conventional 80% power at an α-level of 5%[15]. Moreover, we agree that inordinately large sample sizes can sometimes have high cost/yield ratios. Some fields within biomedicine that work with huge datasets with many thousands or even millions of participants and data points are currently experiencing this challenge. However, as we showed in our Analysis article, most neuroscience research is currently on the other end of this scale, with small, underpowered studies dominating. Finally, we do not advocate making decisions and interpreting results as a dichotomous 'success' or 'failure' on the basis of an absolute $p = 0.05$ threshold. However, this is the norm in much current scientific practice, and our Analysis article explored the implications of combining this approach with small *N* studies. Simply changing to another *p*-value threshold does not solve the problem.

What constitutes an appropriate sample size will depend on the magnitude of the effect being sought. In some cases, a small sample will suffice. However, our Analysis article suggests that these are the exception rather than the rule. It is therefore prudent to assume that the average sample size should increase. Incorporating advance

specification of the magnitude of the effect being sought into our theoretical models and analysis plans, and reporting effect sizes and confidence intervals alongside exact p values (rather than $p = \text{NS}$ or $p < 0.05$) would also improve the strength of scientific inference.

Katherine S. Button, Claire Mokrysz and Marcus R. Munafò are at the School of Experimental Psychology, University of Bristol, BS8 1TU, UK.

Katherine S. Button is also at the School of Social and Community Medicine, University of Bristol, BS8 2BN, UK.

John P. A. Ioannidis is at Stanford University School of Medicine, California 94305, USA.

Brian A. Nosek is at the Department of Psychology, University of Virginia and the Center for Open Science, Charlottesville, Virginia 22903, USA.

Jonathan Flint is at the Wellcome Trust Centre for Human Genetics, University of Oxford, OX 7BN, UK.

Emma S. J. Robinson is at the School of Physiology and Pharmacology, University of Bristol, BS8 1TD, UK.

Correspondence to M.R.M.
e-mail: marcus.munafo@bristol.ac.uk

doi:10.1038/nrn3475-c4
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. Quinlan, P. T. Misuse of power: in defence of small-scale science. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c1 (2013).
3. Ashton, J. C. Experimental power comes from powerful theories — the real problem in null hypothesis testing. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c2 (2013).
4. Bacchetti, P. Small sample size is not the real problem. *Nature Rev. Neurosci.* http://dx.doi.org/10.1038/nrn3475-c3 (2013).
5. Friston, K. Ten ironic rules for non-statistical reviewers. *Neuroimage* **61**, 1300–1310 (2012).
6. Ingre, M. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: Comment on Friston (2012). *Neuroimage* http://dx.doi.org/10.1016/j.neuroimage.2013.03.030 (2013).
7. Lindquist, M. A., Caffo, B. & Crainiceanu, C. Ironing out the statistical wrinkles in "ten ironic rules". *Neuroimage* http://dx.doi.org/10.1016/j.neuroimage.2013.02.056 (2013).
8. Friston, K. Sample size and the fallacies of classical inference. *Neuroimage* http://dx.doi.org/10.1016/j.neuroimage.2013.02.057 (2013).
9. Flint, J. & Munafò, M. R. Candidate and non-candidate genes in behavior genetics. *Curr. Opin. Neurobiol.* **23**, 57–61 (2013).
10. Fowler, C. D., Lu, Q., Johnson, P. M., Marks, M. J. & Kenny, P. J. Habenular α5 nicotinic receptor subunit signalling controls nicotine intake. *Nature* **471**, 597–601 (2011).
11. Munafò, M. R. & Flint, J. Dissecting the genetic architecture of human personality. *Trends Cogn. Sci.* **15**, 395–400 (2011).
12. Tressoldi, P. E., Giofre, D., Sella, F. & Cumming, G. High impact = high statistical standards? Not necessarily so. *PLoS ONE* **8**, e56180 (2013).
13. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Persp. Psychol. Sci.* **7**, 615–631 (2012).
14. Inthout, J., Ioannidis, J. P. & Borm, G. F. Obtaining evidence by a single well-powered trial or several modestly powered trials. *Stat. Methods Med. Res.* http://doi:10.1177/0962280212461098 (2012).
15. Ioannidis, J. P., Hozo, I. & Djulbegovic, B. Optimal type I and type II error pairs when the available sample size is fixed. *J. Clin. Epidemiol.* http://dx.doi.org/10.1016/j.jclinepi.2013.03.002 (2013).

Competing interests statement
The authors declare no competing financial interests.