THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2014
http://dx.doi.org/10.1080/17470218.2014.885986

Comment

On the persistence of low power in psychological science

Ivan Vankov[1], Jeffrey Bowers[1], and Marcus R. Munafò[1,2,3]

[1]School of Experimental Psychology, University of Bristol, Bristol, UK
[2]UK Centre for Tobacco and Alcohol Studies, University of Bristol, Bristol, UK
[3]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

Cohen's classic study on statistical power (Cohen, 1962) showed that studies in the 1960 volume of the *Journal of Abnormal and Social Psychology* lacked sufficient power to detect anything other than large effects ($r \sim 0.60$). Sedlmeier and Gigerenzer (Sedlmeier & Gigerenzer, 1989) conducted a similar analysis on studies in the 1984 volume and found that, if anything, the situation had worsened. Recently, Button and colleagues showed that the average power of neuroscience studies is probably around 20% (Button et al., 2013a). Clearly repeated exhortations that researchers should "pay attention to the power of their tests rather than … focus exclusively on the level of significance" (Sedlmeier & Gigerenzer, 1989) have failed. Here we consider why this might be so.

One reason might be a lack of appreciation of the importance of statistical power within a null hypothesis significance testing (NHST) framework. NHST grew out of the distinct statistical theories of Fisher (Fisher, 1955), and Neyman and Pearson (Rucci & Tweney, 1980). From Fisher we take the concept of null hypothesis testing, and from Neyman-Pearson the concepts of Type I (α) and Type II error (β). Power is a concept arising from Neyman-Pearson theory and reflects the likelihood of correctly rejecting the null hypothesis (i.e., $1 - \beta$). However, the hybrid statistical theory typically used leans most heavily on Fisher's concept of null hypothesis testing. Sedlmeier and Gigerenzer (1989) argued that a lack of understanding of these distinctions partly explained the lack of consideration of statistical power: while we (nominally, at least) adhere to a 5% Type I error rate, we pay little attention to the Type II error rate, despite the need to consider *both* when evaluating whether a research finding is likely to be true (Button et al., 2013a).

Another reason might be the incentive structures within which scientists operate. Scientists are human and will therefore respond (consciously or unconsciously) to incentives; when personal success (e.g., promotion) is associated with the quality and (critically) the *quantity* of publications produced, it makes more sense to use finite resources to generate as many publications as possible. A single transformative study in a highly regarded journal might confer the greatest prestige, but this is a high-risk strategy – the experiment may not produce the desired (i.e., publishable) results, or the journal may not accept it for publication

Correspondence should be addressed to Marcus R. Munafò, School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK. E-mail: marcus.munafo@bristol.ac.uk

Funding: JSB gratefully acknowledges funding from the Leverhulme Trust. MRM is a member of the United Kingdom Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged.

(Sekercioglu, 2013). A safer strategy might be to "salami-slice" one's resources to generate more studies, which, with sufficient analytical flexibility (Simmons, Nelson, & Simonsohn, 2011), will almost certainly produce a number of publishable studies (Sullivan, 2007).

There is some support for the second reason. Studies published in some countries may overestimate true effects more than those published in other countries (Fanelli & Ioannidis, 2013; Munafo, Attwood, & Flint, 2008). This may be because, in certain countries, publication in even medium-rank journals confers substantial direct financial rewards on the authors (Shao & Shen, 2011), which may, in turn, be related to overestimates of true effects (Pan, Trikalinos, Kavvoura, Lau, & Ioannidis, 2005). Authors may therefore (consciously or unconsciously) conduct a larger number of smaller studies, which are still likely to generate publishable findings, rather than risk investing their limited resources in a smaller number of larger studies.

However, to the best of our knowledge, the first possible reason has not been systematically explored. We therefore surveyed studies published recently in a high-ranking psychology journal, and contacted authors to establish the rationale used for deciding sample size (see Supplementary Material). This indicated that approximately one third held beliefs that would serve, on average, to reduce statistical power (see Table 1). In particular, they used accepted norms within their area of research to decide on sample size, in the belief that this would be sufficient to replicate previous results (and therefore, presumably, to identify new findings). Given empirical evidence for a high prevalence of findings close to the $p = .05$ threshold (Masicampo & Lalande, 2012), this belief is likely to be unwarranted. If an experiment finds an effect with $p \sim .05$, and we assume the effect size observed is accurate, then if we repeat the experiment with the same sample size we will on average replicate that finding only 50% of the time (see Supplemental Material). In reality, power will be much lower than 50% because the effect size estimate observed in the original estimate is probably an overestimate (Simonsohn, 2013). However, in our survey, over one third of respondents inaccurately believed that in this scenario the finding would replicate over 80% of the time (see Supplemental Material).

There are unlikely to be simple solutions to the continued lack of appreciation of statistical power. One reason for pessimism, as we have shown, is that these concerns are not new: occasional discussion of these issues has not led to any lasting change. Structural change may be required, including more rigorous enforcement by journals and editors of guidelines, which are often found in instructions for authors but not always followed.

Table 1. *Beliefs about sample size and statistical power*

	How did you decide how many persons to test in the first experiment reported in your paper?									
	I used the same sample size as in another study		*I ran a formal power analysis*		*The number is typical for the area*		*Other*		*Total*	
If someone wants to replicate your first study, what sample size would you recommend?	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Half the original sample size	1	1.1	0	0.0	0	0.0	1	1.1	2	2.1
The same sample size	9	9.6	6	6.4	31	33.0	13	13.8	59	62.8
Double the original sample size	1	1.1	2	2.1	6	6.4	5	5.3	14	14.9
Other	3	3.2	1	1.1	4	4.3	11	11.7	19	20.2
Total	14	14.9	9	9.6	41	43.6	30	31.9	94	

Note: Based on 94 responses.

Recently, *Nature* introduced a submission checklist for life sciences articles, which includes a requirement that sample size be justified (http://www.nature.com/authors/policies/checklist.pdf). Other journals are introducing novel submission formats that place greater emphasis on study design (including statistical power) than on results, including Registered Reports at *Cortex*, and Registered Replication Reports at *Perspectives on Psychological Science*.

The poor reproducibility of scientific findings continues to be a cause of major concern. Small studies with low statistical power contribute to this problem (Bertamini & Munafo, 2012; Button et al., 2013a), and arguments in defence of "small-scale science" (Quinlan, 2013) overlook the fact that larger studies protect against inferences from trivial effect sizes by allowing a better estimation of the magnitude of true effects (Button et al., 2013b). Reasons to resist NHST, and in particular the dichotomous interpretation of *p*-values, have been well rehearsed (Sterne & Davey Smith, 2001), and alternative approaches, such as focusing on effect size estimation or implementing Bayesian approaches, do exist. However, while NHST remains the dominant model for statistical inference, we should ensure that it is used appropriately.

Supplemental material

Supplemental material is available via the "Supplemental" tab on the article's online page (http://dx.doi.org/10.1080/17470218.2014.885986).

Original manuscript received 28 October 2013
Accepted revision received 30 October 2013
First published online 4 March 2014

REFERENCES

Bertamini, M., & Munafo, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, *7*(1), 67–71. doi:10.1177/1745691611429353

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013a). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/Nrn3475

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013b). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*, *14*(8), 585–586. doi:10.1038/Nrn3475-C4

Cohen, J. (1962). The Statistical power of abnormal-social psychological-research - a review. *Journal of Abnormal Psychology*, *65*(3), 145–153. doi:10.1037/H0045186

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences USA*.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *17*(1), 69–78.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology, 65*(11), 2271–2279. doi:10.1080/17470218.2012.711335

Munafo, M. R., Attwood, A. S., & Flint, J. (2008). Bias in genetic association studies: Effects of research location and resources. *Psychological Medicine, 38*(8), 1213–1214. doi:10.1017/S003329170800353x

Pan, Z. L., Trikalinos, T. A., Kavvoura, F. K., Lau, J., & Ioannidis, J. P. A. (2005). Local literature bias in genetic epidemiology: An empirical evaluation of the Chinese literature. *PLOS Medicine*, *2*(12), 1309–1317. doi:10.1371/journal.pmed.0020334

Quinlan, P. T. (2013). Misuse of power: In defence of small-scale science. *Nature Reviews Neuroscience*, *14*(8), 585. doi:10.1038/Nrn3475-C1

Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the 2nd discipline of scientific psychology - historical account. *Psychological Bulletin*, *87*(1), 166–184. doi:10.1037/0033-2909.87.1.166

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, *105*(2), 309–316. doi:10.1037//0033-2909.105.2.309

Sekercioglu, C. H. (2013). Citation opportunity cost of the high impact factor obsession. *Current Biology*, *23*(17), R701–R702.

Shao, J. F., & Shen, H. Y. (2011). The outflow of academic papers from China: Why is it happening and can it be stemmed? *Learned Publishing*, *24*(2), 95–97. doi:10.1087/20110203

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in

data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. doi:10.1177/0956797611417632

Simonsohn, U. (2013). The folly of powering replications based on observed effect size. Retrieved from http://datacolada.org/2013/10/14/powering_replications/

Sterne, J. A., & Davey Smith, G. (2001). Sifting the evidence-what's wrong with significance tests? *British Medical Journal*, *322*(7280), 226–231.

Sullivan, P. F. (2007). Spurious genetic associations. *Biological Psychiatry, 61*(10), 1121–1126. doi:10.1016/j.biopsych.2006.11.010