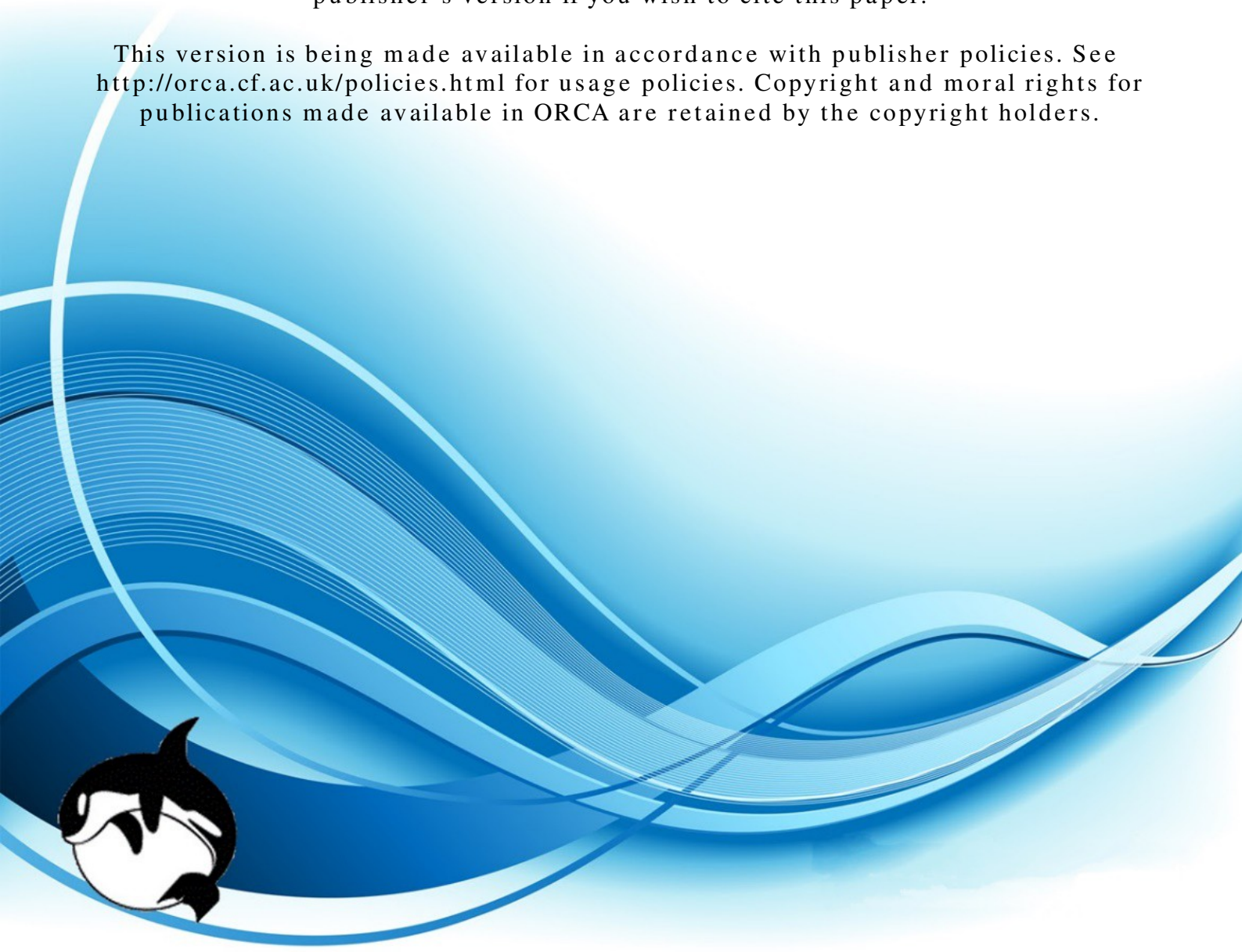

May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them?

Review of **Nord CL**, **Valton V**, **Wood J**, **Roiser JP**. Power-up: a reanalysis of "power failure" in neuroscience using mixture modelling. *J. Neurosci* 37 (34): 3592-16, 2017.

Johannes Algermissen[1§], David M. A. Mehler[2§*]

[1] Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

[2] Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, United Kingdom

* Correspondence: David M. A. Mehler, Cardiff University Brain Research Imaging Centre (CUBRIC), Maindy Road, Cardiff CF24 4HQ, United Kingdom, E-mail: MehlerD@cardiff.ac.uk

Running head: May the power be with you

[§] Both authors contributed equally to this work.

Abstract

Statistical power is essential for robust science and replicability, but a meta-analysis by Button et al. in 2013 diagnosed a "power failure" for neuroscience. In contrast, Nord et al. (*J Neurosci* 37: 8051-8061, 2017) re-analyzed these data and suggested that some studies feature high power. We illustrate how publication and researcher bias might have inflated power estimates, and review recently introduced techniques that can improve analysis pipelines and increase power in neuroscience studies.

Keywords

statistical power; meta-analysis; neuroscience; bias

Many scientific disciplines, including psychology, medicine, and neuroscience currently suffer from *low statistical power*, i.e. they have a low chance to detect the effects they investigate. One of the main reasons for low power are small sample sizes. These usually contain higher levels of noise and are thus less likely to find an effect. However, if a statistically significant result is found with a small sample, some researchers tend to believe that such results must reflect a truly large effect ("what does not kill my effect makes it stronger"; Loken and Gelman, 2017). This belief is misleading because the increased noise in small studies makes effect size estimates imprecise and increases their variability (see also shape of distributions in **Figure 1**). In fact, significant estimates are often *inflated*, i.e. much larger than the *true* effect size (Loken and Gelman 2017). Recent estimates suggest that for this reason, more than 50% of published findings in neuroscience are likely to be *false positives* (Szucs and Ioannidis 2017): treatments that are reported to work may not work reliably, genes that are reported to contribute to a phenotype may contribute little, and conditions that are reported to matter for cognitive processes may only play a marginal role.

What are the underlying reasons for the high rate of false positives in science articles? *Publication bias* is one main reason: significant results are more likely to be accepted for publication than nonsignificant results (Dwan et al. 2008). Another reason for the high rate of false positive findings is *researcher bias*: questionable research practices — such as generating hypotheses after looking at the data, selecting dependent and control variables post-hoc, defining data exclusion criteria post-hoc, and reporting results selectively based on their statistical outcome — can increase the likelihood of *false positive* results (Munafò et al. 2017). Furthermore, fields that work with high dimensional data, such as produced by brain signals, require complex "analysis pipelines". These usually involve numerous pre-processing and data analysis steps, which often result in many ways to analyze such data. In

consequence, different analysis pipelines can lead to vastly different analysis outcomes and interpretations (Carp 2012).

Questionable research practices have been investigated for different neuroscience fields. For functional neuroimaging, Carp (2012) demonstrated how exhaustive combinations of possible pre-processing and data analysis steps results in several thousand unique analysis pipelines. Their results varied remarkably with regards to brain activation strength, location, and extent. For event-related potentials (ERPs) in electrophysiology, Luck and Gaspelin (2017) demonstrated how the common practice of first selecting time windows based on a test statistic (e.g. the grand average) and then comparing conditions on the very same statistic may yield statistically significant, but hardly replicable results. For non-invasive brain stimulation, Héroux et al. (2017) investigated the prevalence of questionable research practices among researchers who work with brain stimulation techniques. In their survey, the authors found that a high proportion admitted to committing questionable research practices such as selective reporting of outcomes and adjusting statistical analyses to reach significant results. As we would expect, when researchers tweak analyses to reach significant results, small or non-existent effects become inflated and appear more reliable in the literature than they really are.

To counter questionable research practices and improve replicability, funders and publishers increasingly urge researchers to adopt more rigorous research practices, including pre-registrations and *a-priori* power calculations (Munafò et al. 2017). These calls seem timely given that in 2013, Button et al.'s seminal meta-analysis diagnosed a “power failure” in neuroscience. However, one remaining question was whether low power affected all of neuroscience, or only certain subfields.

In a study recently published in *The Journal of Neuroscience*, Nord et al. (2017) re-analyzed Button et al.'s (2013) data to test whether their sample contained distinct subsets of studies with different degrees of statistical power. Button et al. reported an alarmingly low median power of only 0.21, which means that only once in five times, studies could detect the effect they were investigating. Button et al. performed a "meta-meta-analysis" on all meta-analyses published in neuroscience in 2011 ($N = 49$), assuming that all studies stemmed from the same population of studies. However, while most studies had very low statistical power, the descriptive statistics in Button et al. suggested that a small proportion of studies had very high power (*Figure 3* in Button et al.). In response, Nord et al. proposed that these studies likely stemmed from different underlying subpopulations of studies, i.e. the data were heterogeneous. Nord et al. tested this proposition using Gaussian mixture modelling (GMM), a technique that fits a pre-specified number of separate normal distributions to an observed distribution. For heterogeneous data, this method is more informative than a single summary statistic (such as the median) because GMM can cope with multimodal distributions. For instance, if a data set featured many low and a few highly powered studies, a median merely reports that (at least) 50% of these studies feature low power. In contrast, GMM can infer that a distinct subset of highly powered studies exists and hence allows a more nuanced interpretation of the data. Nord et al. estimated the power of each single study ($N = 730$) based on their sample size and their weighted mean effect size (as reported in the respective original meta-analysis). They fitted models with different numbers of underlying normal distributions and determined which model fitted the data best (*Figure 2* in Nord et al.).

Nord et al. indeed found indicators for highly powered studies, thereby challenging Button et al.'s conclusion that there is a general "power failure" in neuroscience. Foremost, the data were best described by four underlying normal distributions, one of which covered studies with very high power. Hence, if interpreted as a single representative number, the median

power of 0.21 reported by Button et al. was misleading. In fact, over 70% of studies featured power of less than 0.5 (i.e. less than the chance level of landing heads or tails in a coin toss). However, their data also suggested that ~13% of studies appeared sufficiently or even highly powered (> 0.80; *Figure 3a* in Nord et al.). Moreover, Nord et al. pointed out that in total, seven meta-analyses found null results. If an effect does not exist, it cannot be detected, and power is hence not defined. After excluding studies that reported null results, the median power increased to 0.30. Lastly, the authors investigated the composition of power distributions for the subfields of genetics, psychology, neuroimaging, treatment, neurochemistry, and miscellaneous, separately. Notably, these fields work with very different data types and effect sizes. They found that gene association studies in particular, which composed one third of the sample, featured mainly very low-powered (<0.2) studies. It should be noted, however, that this field has formed large consortia to increase power, for instance ENIGMA and CommonMind[1]. Hence, statistical power for more recently published gene association studies has likely improved.

Taken together, Nord et al. seemed to extend Button et al.'s finding, showing that power in their data set was heterogenous. However, Nord et al.'s analyses were limited by the data because they included exclusively published studies, which likely reported inflated power estimates due to publication bias. High power estimates can occur with a) large samples that can detect small, moderate, and large effect sizes, and b) small samples that can only pick up large effect sizes—which are likely inflated estimates of small effects. The probability that a reported power estimate reflects truly high power (case a) can be inferred from three assumptions (Szucs and Ioannidis, 2017): 1) Only few effects are truly large, but many are

[1] ENIGMA (Enhancing Neuro Imaging Genetics Through Meta Analysis) is a network of researchers in neuroscience imaging genomics http://enigma.ini.usc.edu/). CommonMind is a public-private partnership that pursues projects within and outside of neuroscience (http://sagebase.org/research-projects/the-commonmind-consortium/).

small; 2) in typical, small samples, small effects can only become significant if they are inflated (Loken and Gelman 2017); and 3) significant effects are more likely to be published (publication bias). This effect is also illustrated in **Figure 1**. Small sample sizes result in larger variability and hence a broader distribution (see red distribution) compared to large sample sizes (see green distribution). Assuming publication bias and a true effect size of d = 0.30 (often considered a moderate effect size), small studies with significant results overestimate the true effect more than large studies with significant results. Therefore, among the studies published and included in meta-analyses, there will be more studies that overestimate effect sizes—and hence create the illusion of high power—than studies that estimate effect sizes accurately.

Altogether, in the presence of publication and researcher bias, large reported effects (and power estimates) are more likely to reflect small effects that are inflated than truly large effects. Therefore, such biases cannot only distort the estimates of single studies but might even lead to overestimations in meta-analyses. Crucially, both Button et al. as well as Nord et al. focused on sample size as the sole determinant of power. However, besides using larger samples, choosing more efficient analysis techniques can also increase power. In the following paragraphs, we will review recent developments in model-based (multilevel models) and model-free (machine learning) approaches that allow for a more efficient data usage.

How can neuroscientists solve their power problem? First, they can improve their power calculations. Researchers should calculate power *before* data collection and specify their *smallest effect size of interest* (SESOI; Lakens et al. 2018). They should neither rely on effect sizes reported in the literature, which are often inflated, nor on effect size estimates from small-sample pilot studies, which vary largely (**Figure 1**, red distribution) and might thus

177 severely underestimate the sample size required for adequate power. In contrast, SESOIs
178 require that researchers specify the smallest effect size they consider worthwhile
179 investigating. SESOIs may vary between different fields and hypotheses. For instance,
180 translational researchers may use minimal clinically important differences (MCIDs) for an
181 outcome variable to power intervention studies. Taken together, researchers who work with
182 SESOIs are more likely to conduct adequately powered studies.
183
184 In addition to using larger sample sizes, researchers can also employ repeated-measures
185 designs to increase power, e.g. by collecting multiple measures of the same individual and
186 analyzing data with multilevel models (also called "hierarchical models" or "mixed effects
187 models"; Aarts et al. 2015): Often, experiments yield so-called *nested* data, e.g., recordings of
188 multiple trials performed by the same subject or nerve cells from the same cell colony. Data
189 points from the same source are on average more similar than data points from different
190 sources. Hence, the error terms of data points from the same source are correlated, and the
191 assumption of *independent* observations is violated. Traditional approaches account for this
192 structure by aggregating across trials and performing statistical tests on the *average* responses
193 of subjects. However, these approaches reduce meaningful within-subject variance, which
194 decreases power and makes tests more susceptible to unbalanced designs, missing data, and
195 outliers. In contrast, multilevel models can fit the effects of experimental manipulations for
196 each subject separately (*random effects)*, as well as for the entire sample (*fixed effects)*. By
197 “shrinking” estimates of individual subjects to values closer to the group-level mean (Aarts et
198 al. 2015), multilevel models decrease the influence of outliers and account for regression to
199 the mean, resulting in more robust estimates. Thereby, the use of multilevel models can
200 decrease the rates of false positive findings and increase replicability.
201

Lastly, with noisy measurements, observed effects are likely to be small, but more efficient pipelines can increase power. For instance, novel real-time optimisation techniques can increase the quality of neuroimaging recordings as well as effect sizes in cognitive or behavioural tasks during data acquisition. A recently introduced machine learning technique enables algorithms to learn a stimulus-brain response relationship and adaptively choose stimuli or conditions based on the subject's individual brain responses ("Neuroadaptive Bayesian optimisation"; Lorenz et al. 2017). Researchers may for example investigate which cognitive tasks can optimally disambiguate activity between overlapping, yet distinct brain networks. The algorithm will explore a given set of experimental paradigms and learn which stimuli can best disambiguate between the networks. In a similar way, real-time optimisation can be applied in other contexts to yield more efficient experimental parameters. For instance, in brain stimulation studies, an optimisation algorithm can learn which subject-specific frequency and intensity settings yield large brain responses (Lorenz et al. 2017). Moreover, real-time optimisation can help to fulfil pre-specified data quality standards. For instance, head motion can corrupt fMRI data, however, real-time optimisation algorithms can flexibly adapt sequences to minimize the proportion of images with inacceptable head-motion. Taken together, real-time applications allow researchers to optimise their parameters of interest and minimise the impact of noise. Lastly, since real-time experiments require that researchers specify the search space and parameters in advance, they can effectively reduce researcher bias.

In conclusion, Nord and colleagues have complemented Button et al. by demonstrating how to detect heterogeneity in meta-analytic data. They have suggested that some neuroscience studies may be highly powered. However, this NeuroForum article argues that high power estimates found in the current literature are more likely to stem from overestimations of small effects—driven by publication and researcher bias—than from truly adequately powered

studies. We have presented three approaches that can help neuroscientists to improve power without increasing sample size. Once researchers specify SESOIs for adequate power analyses, use more efficient analysis techniques, and pre-register their hypotheses and analyses, published effect size and power estimates will become more credible and the literature less biased. Future neuroscience meta-analyses could benefit from Gaussian mixture modelling as used by Nord et al., for example when monitoring how the above-mentioned developments impact replicability in neuroscience. As this technique can detect differences within a set of studies, it may help identify the factors that are most effective in increasing power.

Acknowledgements

We thank Dr. Chris Chambers, Dr. Paul Hanel, Robert Thibault, and Dr. Rhian Barrance for helpful comments on a draft of this article.

Grants

Johannes Algermissen is supported by a PhD studentship from Radboud University. David M.A. Mehler is supported by PhD studentship from Health and Care Research Wales (HS/14/20).

Disclosures

The authors declare no conflicts of interest.

References

Aarts E, **Dolan C V**, **Verhage M**, **van der Sluis S**. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neurosci* 16: 94, 2015.

Button KS, **Ioannidis JPA**, **Mokrysz C**, **Nosek BA**, **Flint J**, **Robinson ESJ**, **Munafò MR**. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14: 365–376, 2013.

Carp J. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Front Neurosci* 6: 149, 2012.

Dwan K, **Altman DG**, **Arnaiz JA**, **Bloom J**, **Chan A-W**, **Cronin E**, **Decullier E**, **Easterbrook PJ**, **Von Elm E**, **Gamble C**, **Ghersi D**, **Ioannidis JPA**, **Simes J**, **Williamson PR**. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 3: e3081, 2008.

Héroux ME, **Loo CK**, **Taylor JL**, **Gandevia SC**. Questionable science and reproducibility in electrical brain stimulation research. *PLoS One* 12: e0175635, 2017.

Lakens D, **Scheel AM**, **Isager P**. Equivalence testing for psychological research: A tutorial. *Adv. Methods Pract. Psychol. Sci.* (2018). doi: 10.1007/s11947-009-0181-3.

Loken E, **Gelman A**. Measurement error and the replication crisis. *Science (80-)* 355: 584–585, 2017.

Lorenz R, **Hampshire A**, **Leech R**. Neuroadaptive Bayesian optimization and hypothesis testing. *Trends Cogn Sci* 21: 155–167, 2017.

Luck SJ, **Gaspelin N**. How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology* 54: 146–157, 2017.

Munafò MR, **Nosek BA**, **Bishop DVM**, **Button KS**, **Chambers CD**, **Percie du Sert N**, **Simonsohn U**, **Wagenmakers E-J**, **Ware JJ**, **Ioannidis JPA**. A manifesto for reproducible science. *Nat Hum Behav* 1: 21, 2017.

Nord CL, **Valton V**, **Wood J**, **Roiser JP**. Power-up: a reanalysis of "power failure" in neuroscience using mixture modelling. *J Neurosci* 37: 3592–16, 2017.

Szucs D, **Ioannidis JPA**. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biol* 15: e2000797, 2017.

Figure captions

Figure 1: Distribution of sample estimates of a small effect either in large studies (green distribution) or in small studies (red distribution). Shaded areas indicate reported effects if only significant results are reported (publication bias). When there is a true effect of $d = 0.30$ (cyan vertical line), most studies (80%) with large samples will detect it and yield a significant result for effect size estimates > 0.21 (shaded in green). In contrast, studies with small samples can only detect it for effect size estimates > 0.42, and thus only a small fraction (30%) will detect the effect (shaded in red). In the presence of strong publication bias, small-sample studies only get published when they yield a significant result. Such studies will always overestimate the true effect (indicated by the lack of an overlap between the red shaded area and the cyan vertical line) and will do so to a greater extent than large published studies (see difference between green and red vertical line). The following parameters were used to create the figure: The small sample size ($N = 25$) is based on 0.30 power to detect an effect of Cohen's d = 0.30. Power of 0.30 is equivalent to the median power in neuroscience found by Nord et al. after excluding null results from meta-analyses. The large sample size ($N = 90$) is based on a hypothetical statistical power of 0.80, which is a value that is often recommended. Shown are results for a one-sample two-sided t-test at an alpha level of 0.05.

Overestimation of Small Effects Given Publication Bias