

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <http://orca.cf.ac.uk/109973/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Algermissen, Johannes and Mehler, David 2018. May the power be with you: are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology* 119 (6) , pp. 2114-2117. 10.1152/jn.00765.2017 file

Publishers page: <http://dx.doi.org/10.1152/jn.00765.2017> <<http://dx.doi.org/10.1152/jn.00765.2017>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



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

3

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

6

7 Johannes Algermissen^{1§}, David M. A. Mehler^{2§*}

8

9 ¹ Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The
10 Netherlands

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

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

16

17 Running head: May the power be with you

18

19 [§] Both authors contributed equally to this work.

20

21

22

23

24

25

26

27
28
29
30
31
32
33

34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

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

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

65

66 What are the underlying reasons for the high rate of false positives in science articles?
67 *Publication bias* is one main reason: significant results are more likely to be accepted for
68 publication than nonsignificant results (Dwan et al. 2008). Another reason for the high rate of
69 false positive findings is *researcher bias*: questionable research practices — such as
70 generating hypotheses after looking at the data, selecting dependent and control variables
71 post-hoc, defining data exclusion criteria post-hoc, and reporting results selectively based on
72 their statistical outcome — can increase the likelihood of *false positive* results (Munafò et al.
73 2017). Furthermore, fields that work with high dimensional data, such as produced by brain
74 signals, require complex “analysis pipelines”. These usually involve numerous pre-processing
75 and data analysis steps, which often result in many ways to analyze such data. In

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

78

79 Questionable research practices have been investigated for different neuroscience fields. For
80 functional neuroimaging, Carp (2012) demonstrated how exhaustive combinations of possible
81 pre-processing and data analysis steps results in several thousand unique analysis pipelines.

82 Their results varied remarkably with regards to brain activation strength, location, and extent.

83 For event-related potentials (ERPs) in electrophysiology, Luck and Gaspelin (2017)

84 demonstrated how the common practice of first selecting time windows based on a test

85 statistic (e.g. the grand average) and then comparing conditions on the very same statistic may

86 yield statistically significant, but hardly replicable results. For non-invasive brain stimulation,

87 Héroux et al. (2017) investigated the prevalence of questionable research practices among

88 researchers who work with brain stimulation techniques. In their survey, the authors found

89 that a high proportion admitted to committing questionable research practices such as

90 selective reporting of outcomes and adjusting statistical analyses to reach significant results.

91 As we would expect, when researchers tweak analyses to reach significant results, small or

92 non-existent effects become inflated and appear more reliable in the literature than they really

93 are.

94

95 To counter questionable research practices and improve replicability, funders and publishers

96 increasingly urge researchers to adopt more rigorous research practices, including pre-

97 registrations and *a-priori* power calculations (Munafò et al. 2017). These calls seem timely

98 given that in 2013, Button et al.'s seminal meta-analysis diagnosed a “power failure” in

99 neuroscience. However, one remaining question was whether low power affected all of

100 neuroscience, or only certain subfields.

101

102 In a study recently published in *The Journal of Neuroscience*, Nord et al. (2017) re-analyzed
103 Button et al.'s (2013) data to test whether their sample contained distinct subsets of studies
104 with different degrees of statistical power. Button et al. reported an alarmingly low median
105 power of only 0.21, which means that only once in five times, studies could detect the effect
106 they were investigating. Button et al. performed a “meta-meta-analysis” on all meta-analyses
107 published in neuroscience in 2011 ($N = 49$), assuming that all studies stemmed from the same
108 population of studies. However, while most studies had very low statistical power, the
109 descriptive statistics in Button et al. suggested that a small proportion of studies had very high
110 power (*Figure 3* in Button et al.). In response, Nord et al. proposed that these studies likely
111 stemmed from different underlying subpopulations of studies, i.e. the data were
112 heterogeneous. Nord et al. tested this proposition using Gaussian mixture modelling (GMM),
113 a technique that fits a pre-specified number of separate normal distributions to an observed
114 distribution. For heterogeneous data, this method is more informative than a single summary
115 statistic (such as the median) because GMM can cope with multimodal distributions. For
116 instance, if a data set featured many low and a few highly powered studies, a median merely
117 reports that (at least) 50% of these studies feature low power. In contrast, GMM can infer that
118 a distinct subset of highly powered studies exists and hence allows a more nuanced
119 interpretation of the data. Nord et al. estimated the power of each single study ($N = 730$)
120 based on their sample size and their weighted mean effect size (as reported in the respective
121 original meta-analysis). They fitted models with different numbers of underlying normal
122 distributions and determined which model fitted the data best (*Figure 2* in Nord et al.).
123
124 Nord et al. indeed found indicators for highly powered studies, thereby challenging Button et
125 al.'s conclusion that there is a general “power failure” in neuroscience. Foremost, the data
126 were best described by four underlying normal distributions, one of which covered studies
127 with very high power. Hence, if interpreted as a single representative number, the median

128 power of 0.21 reported by Button et al. was misleading. In fact, over 70% of studies featured
129 power of less than 0.5 (i.e. less than the chance level of landing heads or tails in a coin toss).
130 However, their data also suggested that ~13% of studies appeared sufficiently or even highly
131 powered (> 0.80 ; *Figure 3a* in Nord et al.). Moreover, Nord et al. pointed out that in total,
132 seven meta-analyses found null results. If an effect does not exist, it cannot be detected, and
133 power is hence not defined. After excluding studies that reported null results, the median
134 power increased to 0.30. Lastly, the authors investigated the composition of power
135 distributions for the subfields of genetics, psychology, neuroimaging, treatment,
136 neurochemistry, and miscellaneous, separately. Notably, these fields work with very different
137 data types and effect sizes. They found that gene association studies in particular, which
138 composed one third of the sample, featured mainly very low-powered (<0.2) studies. It should
139 be noted, however, that this field has formed large consortia to increase power, for instance
140 ENIGMA and CommonMind¹. Hence, statistical power for more recently published gene
141 association studies has likely improved.

142

143 Taken together, Nord et al. seemed to extend Button et al.'s finding, showing that power in
144 their data set was heterogenous. However, Nord et al.'s analyses were limited by the data
145 because they included exclusively published studies, which likely reported inflated power
146 estimates due to publication bias. High power estimates can occur with a) large samples that
147 can detect small, moderate, and large effect sizes, and b) small samples that can only pick up
148 large effect sizes—which are likely inflated estimates of small effects. The probability that a
149 reported power estimate reflects truly high power (case a) can be inferred from three
150 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/>).

151 small; 2) in typical, small samples, small effects can only become significant if they are
152 inflated (Loken and Gelman 2017); and 3) significant effects are more likely to be published
153 (publication bias). This effect is also illustrated in **Figure 1**. Small sample sizes result in
154 larger variability and hence a broader distribution (see red distribution) compared to large
155 sample sizes (see green distribution). Assuming publication bias and a true effect size of $d =$
156 0.30 (often considered a moderate effect size), small studies with significant results
157 overestimate the true effect more than large studies with significant results. Therefore, among
158 the studies published and included in meta-analyses, there will be more studies that
159 overestimate effect sizes—and hence create the illusion of high power—than studies that
160 estimate effect sizes accurately.

161

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

171

172 How can neuroscientists solve their power problem? First, they can improve their power
173 calculations. Researchers should calculate power *before* data collection and specify their
174 *smallest effect size of interest* (SESOI; Lakens et al. 2018). They should neither rely on effect
175 sizes reported in the literature, which are often inflated, nor on effect size estimates from
176 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

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

222

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

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

237

238

239

240

241 **Acknowledgements**

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

244 **Grants**

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

248 **Disclosures**

249 The authors declare no conflicts of interest.

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267 **References**

- 268 **Aarts E, Dolan C V, Verhage M, van der Sluis S.** Multilevel analysis quantifies variation in
269 the experimental effect while optimizing power and preventing false positives. *BMC Neurosci*
270 16: 94, 2015.
- 271 **Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR.**
272 Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev*
273 *Neurosci* 14: 365–376, 2013.
- 274 **Carp J.** On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of
275 fMRI Experiments. *Front Neurosci* 6: 149, 2012.
- 276 **Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, Cronin E, Decullier E,**
277 **Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JPA, Simes J, Williamson**
278 **PR.** Systematic review of the empirical evidence of study publication bias and outcome
279 reporting bias. *PLoS One* 3: e3081, 2008.
- 280 **Héroux ME, Loo CK, Taylor JL, Gandevia SC.** Questionable science and reproducibility
281 in electrical brain stimulation research. *PLoS One* 12: e0175635, 2017.
- 282 **Lakens D, Scheel AM, Isager P.** Equivalence testing for psychological research: A tutorial.
283 *Adv. Methods Pract. Psychol. Sci.* (2018). doi: 10.1007/s11947-009-0181-3.
- 284 **Loken E, Gelman A.** Measurement error and the replication crisis. *Science* (80-) 355: 584–
285 585, 2017.
- 286 **Lorenz R, Hampshire A, Leech R.** Neuroadaptive Bayesian optimization and hypothesis
287 testing. *Trends Cogn Sci* 21: 155–167, 2017.
- 288 **Luck SJ, Gaspelin N.** How to get statistically significant effects in any ERP experiment (and
289 why you shouldn't). *Psychophysiology* 54: 146–157, 2017.
- 290 **Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N,**
291 **Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA.** A manifesto for reproducible
292 science. *Nat Hum Behav* 1: 21, 2017.

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

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

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316 **Figure captions**

317 **Figure 1:** Distribution of sample estimates of a small effect either in large studies (green
318 distribution) or in small studies (red distribution). Shaded areas indicate reported effects if
319 only significant results are reported (publication bias). When there is a true effect of $d = 0.30$
320 (cyan vertical line), most studies (80%) with large samples will detect it and yield a
321 significant result for effect size estimates > 0.21 (shaded in green). In contrast, studies with
322 small samples can only detect it for effect size estimates > 0.42 , and thus only a small fraction
323 (30%) will detect the effect (shaded in red). In the presence of strong publication bias, small-
324 sample studies only get published when they yield a significant result. Such studies will
325 always overestimate the true effect (indicated by the lack of an overlap between the red
326 shaded area and the cyan vertical line) and will do so to a greater extent than large published
327 studies (see difference between green and red vertical line). The following parameters were
328 used to create the figure: The small sample size ($N = 25$) is based on 0.30 power to detect an
329 effect of Cohen's $d = 0.30$. Power of 0.30 is equivalent to the median power in neuroscience
330 found by Nord et al. after excluding null results from meta-analyses. The large sample size (N
331 $= 90$) is based on a hypothetical statistical power of 0.80, which is a value that is often
332 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

