NeuroImage: Clinical 35 (2022) 103008

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Excess significance and power miscalculations in neurofeedback research

Robert T. Thibault [a,b,*,1], Hugo Pedder [c,2]

[a] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, United States
[b] School of Psychological Science, University of Bristol, United Kingdom
[c] Department Population Health Sciences, Bristol Medical School, University of Bristol, United Kingdom

ARTICLE INFO

Keywords:
Neurofeedback
Neuroimaging
fMRI
fNIRS
Statistical power analysis

Recent systematic reviews of neurofeedback with functional magnetic resonance imaging (fMRI-nf) (Tursic et al., 2020) and neurofeedback with functional near infrared spectroscopy (fNIRS-nf) (Kohl et al., 2020) miscalculate the statistical power and statistical sensitivity of several studies they review. The fMRI-nf review overestimates the mean and median statistical power of included studies by about 3 times and the statistical sensitivity by about 2 times (see Table 1 for recalculated values and comparisons). The fNIRS-nf review, on which I (RTT) was a coauthor, overestimates power by about 2 times and sensitivity by about 1.5 times (see Table 2).

The miscalculations arise from an easy-to-miss default option for repeated measures (mixed)[3] ANOVAs in the statistical software program GPower (Faul et al., 2007), which both reviews used (see Fig. 1 for a depiction)[4]. The default option defines a variable in the effect size calculation (η^2_p) in such a way that the common usage of small, medium, and large effects sizes for the interaction of repeated measures (mixed) ANOVAs (*f*) doesn't hold true. If unaware of the default option, the power calculations will account for the correlations between repeated measures a second time, and in turn substantially—but erroneously—increase power. The GPower software itself highlights that Cohen (1988) recommended another option (as viewable in Fig. 1). While Lakens (2013) explained this issue almost 10 years ago, it remains likely that researchers continue to use GPower without awareness of this default option and its implications[5]. Fortunately, the authors of both reviews published their data as supplementary material, making reanalysis possible.

We recalculated the statistical power and sensitivity of the studies from Tursic et al., 2020 and Kohl et al., 2020 using the WebPower package in R[6]. Our recalculations show that the median study in the fMRI-nf review has only 21% power to detect clinical effects of Cohen's

* Corresponding author at: Meta-Research Innovation Center at Stanford (METRICS), Stanford University, United States.
E-mail address: robert.thibault@stanford.edu (R.T. Thibault).
[1] ORCID: 0000-0002-6561-3962.
[2] ORCID: 0000-0002-7813-3749.

[3] We use the term "repeated measures (mixed) ANOVA" to describe a study design with one independent measure and one repeated measure (e.g., two independent groups measured at two time points). GPower uses the term "ANOVA: Repeated measures" for this type of design.

[4] Kohl and coauthors (including me—RTT) were unaware of this default option in GPower. I was only led to become aware of this issue when reading the fMRI-nf review and noticing a surprising amount of statistical power.

[5] Kieslich (2020) provides a detailed explanation of how GPower calculates *f* for the different options for repeated measures (mixed) ANOVAs. The GPower team informed us that Cohen proposed Cohen's *f* for between-subjects designs and that this is how GPower defines *f* when using the default option, regardless of the ANOVA design the user selects.

[6] We performed our calculations in R. They are all available in our open code. We reproduce Table 3 from Tursic et al. and Table 5 from Kohl et al. using the sample sizes and statistical tests which each review provides in their supplementary material (i.e., we did not re-extract this information from the original studies or check if the statistical tests in the original studies were appropriate). The WebPower package we used assumes a correlation of 0.7 for repeated measures for ANOVAs, which is slightly more conservative than the 0.8 correlation used in the reviews.

https://doi.org/10.1016/j.nicl.2022.103008
Received 11 April 2022; Accepted 12 April 2022
Available online 26 April 2022

Table 1
Recalculated values for the fMRI-nf review (Tursic et al., 2020).

		Power			Sensitivity (in Cohen's *d*)	
	N	d = 0.2	d = 0.5	d = 0.8	Power = 80%	Power = 95%
Recalculated						
Mean (regulation)	29.22	0.08	0.25	0.47	1.31	1.68
Median (regulation)	22.00	0.07	0.20	0.43	1.26	1.62
Mean (clinical)	26.73	0.07	0.21	0.45	1.31	1.68
Median (clinical)	27.00	0.07	0.21	0.48	1.15	1.47
Original						
Mean (regulation)	29.90	0.24	0.61	0.76	0.77	0.99
Median (regulation)	22.50	0.15	0.67	0.98	0.58	0.73
Mean (clinical)	26.70	0.31	0.73	0.85	0.58	0.74
Median (clinical)	27.00	0.30	0.98	0.99	0.36	0.46
Overestimation factor (original/recalculated)						
Mean (regulation)	1.02	2.91	2.49	1.61	1.70	1.69
Median (regulation)	1.02	2.05	3.34	2.27	2.17	2.22
Mean (clinical)	1.00	4.20	3.48	1.89	2.26	2.27
Median (clinical)	1.00	4.10	4.77	2.05	3.18	3.21

The first section of the table presents the values we calculated. The second section presents the values published in the original review. The third section presents an overestimation factor calculated by dividing the original values by the recalculated values for power and by dividing the recalculated values by the original values for sensitivity. Power and sensitivity calculations for the ability to regulate the neurofeedback signal are presented separately from those for clinical measures. The overestimation factor was calculated before rounding values to two decimal place. Thus, recalculating the overestimation factor with the numbers in the table will produce slightly different values. The mean and median sample sizes in the review differ slightly from ours, possibly due to a calculation error. We used the data provided in the review's supplementary material for these calculations.

Table 2
Recalculated values for the fNIRS-nf review (Kohl et al., 2020).

		Power			Sensitivity (in Cohen's *d*)	
	N	d = 0.2	d = 0.5	d = 0.8	Power = 80%	Power = 95%
Recalculated						
Mean (regulation)	19.29	0.14	0.41	0.67	0.98	1.29
Median (regulation)	19.00	0.14	0.43	0.75	0.85	1.13
Mean (behavioural)	22.10	0.10	0.31	0.56	1.11	1.45
Median (behavioural)	20.00	0.08	0.22	0.42	1.30	1.66
Original						
Mean (regulation)	22.11	0.20	0.55	0.74	0.88	1.15
Median (regulation)	20.00	0.16	0.48	0.80	0.75	1.00
Mean (behavioural)	22.10	0.20	0.68	0.87	0.66	0.87
Median (behavioural)	20.00	0.22	0.76	0.97	0.53	0.69
Overestimation factor (original/recalculated)						
Mean (regulation)	1.15	1.45	1.33	1.10	1.12	1.12
Median (regulation)	1.05	1.14	1.12	1.06	1.14	1.13
Mean (behavioural)	1.00	1.97	2.23	1.55	1.69	1.67
Median (behavioural)	1.00	2.86	3.53	2.30	2.45	2.41

The first section of the table presents the values we calculated. The second section presents the values published in the original review. The third section presents an overestimation factor calculated by dividing the original values by the recalculated values for power and by dividing the recalculated values by the original values for sensitivity. Power and sensitivity calculations for the ability to regulate the neurofeedback signal are presented separately from those for behavioural measures. The overestimation factor was calculated before rounding values to two decimal place. Thus, recalculating the overestimation factor with the numbers in the table will produce slightly different values. The mean and median sample size in the review differ slightly from ours—whereas we calculated these values based on the sample size used in the statistical tests, Kohl et al. calculated them based on the total number of participants. We removed one study from our calculations because it only ran binomial tests within each participant but did not test for group effects. One study used biserial correlation, for which we calculated power as for a Pearson's correlation. One study used an ANCOVA, for which we calculated power using a 2x2 repeated measures (mixed) ANOVA.

$d = 0.5$ and the median study in the fNIRS-nf review has 22% power to detect behavioural effects of the same size. The median studies in fMRI-nf and fNIRS-nf have 80% power to detect large to very large effect sizes ($d = 0.85 – 1.30$)[7] (see the *sensitivity* columns in Tables 1 and 2). The fMRI-nf review overestimates power to a greater degree than the fNIRS-nf review because more of the studies they reviewed used repeated measures (mixed) ANOVAs, where the consequential default option exists.

Effect sizes of this magnitude are uncommon in medicine. When found, they rarely replicate in larger follow up trials (Nagendran et al., 2016). One study compiled meta-analyses of the 20 most common pharmaceutical therapies and found a mean effect size of $d = 0.58$ (median $d = 0.56$) (Leucht et al., 2015). Antidepressants, for example, have an effect size of $d = 0.30$ compared to placebos for treating depression (Cipriani et al., 2018). For a more tangible comparison, the height difference between men and women over the age of 20 in the United States is $d = 1.01$ (National Center for Health Statistics, 2021)[8]. Thus, the median studies in these neurofeedback reviews have 80% power to detect a clinical or behavioural effect size about 4 times larger than antidepressants or slightly larger than the height difference between men and women in the United States.

Given the sample sizes used in the reviewed studies, even if neurofeedback drove "large" clinical or behavioural effects ($d = 0.8$), less than half of studies should have statistically significant results at $p < .05$. And yet, Tursic et al. found that 10/11 (91%) of the fMRI-nf studies that were

[7] Although these reviews calculate power and sensitivity based on the analyses used in each study, an independent sample *t*-test would be sufficient to answer the basic question "do neurofeedback participants improve more than control participants". The median size fMRI-nf study has 80% power to detect an effect size of $d = 1.10$ and the median size fNIRS-nf study $d = 1.32$ for this analysis, with $\alpha = 0.05$. The effect size *f* used for ANOVAs in GPower depends on η^2_p which varies depending on the study design. Given the less intuitive nature of *f* and that a *t*-test directly answers the main question in most trials, t-tests can be preferable.

[8] We performed these calculations based on data provided by the Centers for Disease Control and Prevention on over 10,000 people in the United States.

not pilot studies reported clinical improvements while another review found that 24/35 (69%) of fMRI-nf studies reported behavioural improvements compared to a control group[9] (Thibault et al., 2018). Kohl et al. found that all studies reported improvement in at least one behavioural measure. This excess significance in the fMRI-nf and fNIRS-nf literature may stem from a combination of an absence of corrections for multiple comparisons, data dependent analytical decisions, selective reporting, publication bias, false positives, statistical tests against baseline rather than against a control group, and other sources of bias.

Can we be sure current sample sizes are insufficient? It depends on the question[10]. On the one hand, if the goal is to show that individuals can control their brain imaging data or improve their behaviour *compared to baseline*—where within-sample designs are appropriate and effect sizes may be large—then the upper end of current sample sizes would be sufficient. For example, neurofeedback has driven very large behavioural effects compared to baseline ($\sim d = 1.5$) when using EEG-nf to treat ADHD (Arnold et al., 2021; Schönenberg et al., 2017) or fMRI-nf to treat depression (Mehler et al., 2018; Young et al., 2017). However, these effect sizes are generally much smaller or absent when comparing the experimental group to an active control group (Trambaiolli et al., 2021).

On the other hand, if the goal is to demonstrate that a target neurofeedback protocol outperforms a reasonable control condition or matches the performance of an accepted treatment, then current sample sizes remain inadequate. Continuing to run poorly powered studies fills the literature with noise and wastes resources (Button et al., 2013). Genetics research provides a stark example of this issue. With the advent of inexpensive genome-wide testing (and the associated ability to increase sample sizes by orders of magnitude) the literature on candidate genes was found to be largely noise (Border et al., 2019; Flint & Munafò, 2013).

How should we move forward? Increasing sample size is an obvious, albeit practically challenging, solution. Without an influx of resources, we would need multi-site collaborations (e.g., as done recently for EEG-nf: Arnold et al., 2021). To detect an effect size equivalent to the median effect size for the 20 most common pharmaceuticals would require 102 participants. An effect size equivalent to antidepressants would require 351 participants. These sample sizes can be prohibitive, even for multi-site collaborations. Increasing the effect size presents another option. Neurofeedback publications sometimes identify "responders" and "non-responders" *post hoc*. If these groups could be identified *a priori*, and neurofeedback selectively applied to responders, the group effect would increase. However, repeated efforts to apply this approach in personalized medicine remain largely unsuccessful (Senn, 2018).

Unfortunately, there's no easy solution. In many cases, resources are simply too scarce to answer a research question. We are better off to resist the temptation to forge ahead with uninformative sample sizes, even when incentive structures may encourage us to do so (Higginson & Munafò, 2016). In the words of Doug Altman (1994): "*We need less research, better research, and research done for the right reasons*".

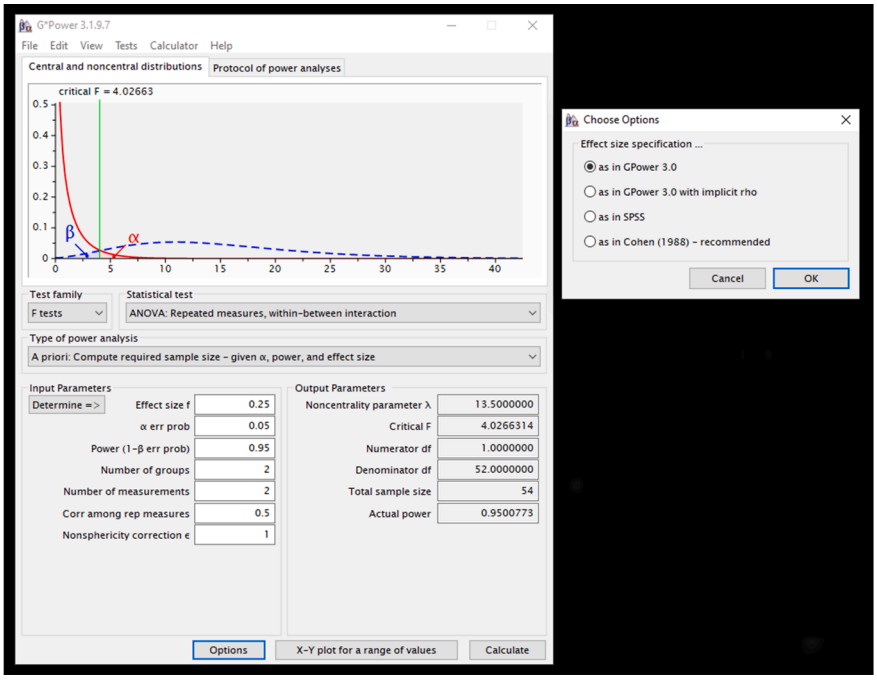

Fig. 1. Depiction of the default and Cohen's recommended options for conducting power calculations for repeated-measures (mixed) ANOVAs in GPower.

[9] The numbers 10/11 (91%) come from data in Table 2 of Tursic et al. for the 11 non-pilot studies included in their power calculations for clinical measures. Their review also presents the numbers: "Out of 78% of studies reporting results, 60% (29/48) reported significant improvement of symptoms". However, the review only calculates power for 27 studies and 16 of these are "pilot, feasibility, or proof-of-pinciple studies…and should therefore also not be performing inferential statistical tests". The number 24/35 (69%) comes from Thibault et al. (2018) Fig 6c, 24 "Yes" and 11 "No".

[10] Our commentary does not comment on the use of neurofeedback for purposes other than regulating brain activity with the aim to impact clinical conditions or behaviour. For example, neurofeedback can be used as a research tool instead of a potential treatment.

Data availability statement

All data and materials related to this study are publicly available on the Stanford Digital Repository (https://doi.org/10.25740/bn925rp5443).

Code availability statement

To facilitate reproducibility this manuscript was written by interleaving regular prose and analysis code using R Markdown. The relevant

files are available on the Stanford Digital Repository (https://doi.org/10.25740/bn925rp5443) and in a Code Ocean container (https://doi.org/10.24433/CO.7282505.v1) which recreates the software environment in which the original analyses were performed. This container allows the manuscript to be reproduced from the data and code with a single button press.

Contributions

Robert Thibault conceived the idea for this commentary and led the analyses and writing. Hugo Pedder provided statistical support, reviewed the code, and contributed to the commentary through discussions.

Funding

Robert Thibault is supported by a general support grant awarded to METRICS from the Laura and John Arnold Foundation and a post-doctoral fellowship from the Fonds de recherche du Québec – Santé. Hugo Pedder was funded by the NIHR Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funders had no role in the data analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

Robert Thibault has received payments for consulting for neurofeedback start-up companies. Hugo Pedder declares no competing interests.

Acknowledgements

We thank the authors of the two reviews for making their data publicly available for reanalysis. We thank Marcus Munafò for discussions on this topic and John Ioannidis for feedback on a draft of this commentary.

References

Altman, D.G., 1994. The scandal of poor medical research. BMJ 308 (6924), 283–284. https://doi.org/10.1136/bmj.308.6924.283.

Arnold, L.E., Arns, M., Barterian, J., Bergman, R., Black, S., Conners, C.K., Connor, S., Dasgupta, S., deBeus, R., Higgins, T., Hirshberg, L., Hollway, J.A., Kerson, C., Lightstone, H., Lofthouse, N., Lubar, J., McBurnett, K., Monastra, V., Buchan-Page, K., Pan, X., Rice, R., Roley-Roberts, M.E., Rhodes, R., Schrader, C., Tan, Y., Williams, C.E., 2021. Double-blind placebo-controlled randomized clinical trial of neurofeedback for attention-deficit/hyperactivity disorder with 13-month follow-up. J. Am. Acad. Child Adolescent Psychiatry 60 (7), 841–855. https://doi.org/10.1016/j.jaac.2020.07.906.

Border, R., Johnson, E.C., Evans, L.M., Smolen, A., Berley, N., Sullivan, P.F., Keller, M.C., 2019. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. Am. J. Psychiatry 176 (5), 376–387. https://doi.org/10.1176/appi.ajp.2018.18070881.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Rev. Neurosci. 14 (5), 365–376. https://doi.org/10.1038/nrn3475.

Cipriani, A., Furukawa, T.A., Salanti, G., Chaimani, A., Atkinson, L.Z., Ogawa, Y., Leucht, S., Ruhe, H.G., Turner, E.H., Higgins, J.P.T., Egger, M., Takeshima, N., Hayasaka, Y.u., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J.P.A., Geddes, J.R., 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet 391 (10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. L. Erlbaum Associates.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191. https://doi.org/10.3758/BF03193146.

Flint, J., Munafò, M.R., 2013. Candidate and non-candidate genes in behavior genetics. Curr. Opin. Neurobiol. 23 (1), 57–61. https://doi.org/10.1016/j.conb.2012.07.005.

Higginson, A.D., Munafò, M.R., 2016. Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLoS Biology 14 (11). https://doi.org/10.1371/journal.pbio.2000995.

Kieslich, P. J., 2020. Cohen's f in repeated measures ANOVAs. https://osf.io/gevp6/.

Kohl, S.H., Mehler, D.M.A., Lührs, M., Thibault, R.T., Konrad, K., Sorger, B., 2020. The potential of functional near-infrared spectroscopy-based neurofeedback—a systematic review and recommendations for best practice. Front. Neurosci. 14 https://doi.org/10.3389/fnins.2020.00594.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4 https://doi.org/10.3389/fpsyg.2013.00863.

Leucht, S., Helfer, B., Gartlehner, G., Davis, J.M., 2015. How effective are common medications: A perspective based on meta-analyses of major drugs. BMC Med. 13 (1), 253. https://doi.org/10.1186/s12916-015-0494-1.

Mehler, D.M.A., Sokunbi, M.O., Habes, I., Barawi, K., Subramanian, L., Range, M., Evans, J., Hood, K., Lührs, M., Keedwell, P., Goebel, R., Linden, D.E.J., 2018. Targeting the affective brain—a randomized controlled trial of real-time fMRI neurofeedback in patients with depression. Neuropsychopharmacology May, 1–8. https://doi.org/10.1038/s41386-018-0126-5.

Nagendran, M., Pereira, T.V., Kiew, G., Altman, D.G., Maruthappu, M., Ioannidis, J.P.A., McCulloch, P., 2016. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ 355, i5432. https://doi.org/10.1136/bmj.i5432.

National Center for Health Statistics, 2021. Vital and Health Statistics, Series 3, Number 46: Anthropometric Reference Data for Children and Adults: United States, 2015–2018. Centers for Disease Control and Prevention, 44.

Schönenberg, M., Wiedemann, E., Schneidt, A., Scheeff, J., Logemann, A., Keune, P.M., Hautzinger, M., 2017. Neurofeedback, sham neurofeedback, and cognitive-behavioural group therapy in adults with attention-deficit hyperactivity disorder: A triple-blind, randomised, controlled trial. Lancet Psychiatry 4 (9), 673–684.

Senn, S., 2018. Statistical pitfalls of personalized medicine. Nature 563 (7733), 619–621. https://doi.org/10.1038/d41586-018-07535-2.

Thibault, R.T., MacPherson, A., Lifshitz, M., Roth, R.R., Raz, A., 2018. Neurofeedback with fMRI: a critical systematic review. NeuroImage 172, 786–807. https://doi.org/10.1016/j.neuroimage.2017.12.071.

Trambaiolli, L.R., Kohl, S.H., Linden, D.E.J., Mehler, D.M.A., 2021. Neurofeedback training in major depressive disorder: a systematic review of clinical efficacy, study quality and reporting practices. Neurosci. Biobehav. Rev. 125, 33–56. https://doi.org/10.1016/j.neubiorev.2021.02.015.

Tursic, A., Eck, J., Lührs, M., Linden, D.E.J., Goebel, R., 2020. A systematic review of fMRI neurofeedback reporting and effects in clinical populations. NeuroImage: Clinical 28, 102496. https://doi.org/10.1016/j.nicl.2020.102496.

Young, K.D., Siegle, G.J., Zotev, V., Phillips, R., Misaki, M., Yuan, H., Drevets, W.C., Bodurka, J., 2017. Randomized clinical trial of real-time fMRI amygdala neurofeedback for major depressive disorder: effects on symptoms and autobiographical memory recall. Am. J. Psychiatry 174 (8), 748–755.