Consideration of Sample Size in Neuroscience Studies

Reproducibility of neuroscience studies is a primary goal of *The Journal of Neuroscience*. There are two main reasons for problems of reproducibility in the neuroscience literature. The first is the inflated false-positive rates that result in many studies falsely rejecting their null hypotheses. This often has its roots in biases in statistical inference. These biases can be introduced by “researcher degrees of freedom,” selecting analytical procedures according to the study outcome; by “hypothesizing after results are known,” offering credibility to tests lacking a hypothesis; or by using parametric procedures when the structure of the data does not warrant them. Such procedural biases and how to minimize them were covered by a previous *JNeurosci* editorial on analytical transparency and reproducibility (Picciotto, 2018).

This editorial focuses on a second reason for limited reproducibility in neuroscience studies: low statistical power, frequently caused by small sample sizes. Here we provide suggestions on how to approach the determination of sample size in the context of the noisy and subtle effects often observed in neuroscience studies. We emphasize how sample size planning depends on whether the statistical goal of the study is to determine the presence of an effect or to obtain accurate estimates of the effect.

Statistical power ($1 - \beta$, where β is the false-negative rate or the probability to fail to reject the null hypothesis when an effect is present) increases with sample size. Given a true effect of a certain size, studies with smaller samples have lower power to detect it. Effects found in studies with low power are subject to the problem of low positive predictive value (Button et al., 2013). On a single test, a small sample size does not inflate the probability of falsely rejecting the null hypothesis (e.g., $\alpha = 5\%$). However, not all researchers are aware that low power in their study increases the probability that their estimation overestimates the true effect size, a situation aptly labeled “Winner’s Curse” (Button et al., 2013).

What does this mean for neuroscience? The real effect sizes for most phenomena uncovered in exploratory studies are in fact smaller than reported, even without accounting for the procedural biases leading to inflated inferential statistics. Therefore, follow-up studies based on those estimated effect sizes should expect to find smaller effects, due to a regression toward the mean of the underlying distribution of effect size. Unfortunately, this phenomenon is often missed given that many studies still fail to report effect size.

These considerations lead us to suggest that, whenever possible, studies should accommodate and plan for two related experiments. First, an “exploratory” experiment provides provisional statistical evidence for the presence of an effect. The findings of this exploratory stage generate (likely inflated) estimates of the effect magnitude, and likely to yield confidence intervals that are wide and imprecise (Maxwell et al., 2008). Second, an “estimation” experiment provides a more precise and accurate estimate of the real sizes of those effects. The exploratory stage could be powered to detect medium to large effect sizes using intermediate sample sizes to avoid the risk of detecting

biologically marginal effects when using large samples (Wilson et al., 2020). The exploratory stage should also quantify the statistical power provided by the experimental design, either *a priori* or with *post hoc* simulations. In contrast, the estimation stage should be used to optimize sample size for effect size estimation. The sample size necessary to obtain an accurate estimate of an effect size is usually larger than the sample size necessary for adequate power to detect the presence of an effect (Maxwell et al., 2008).

Procedurally, this suggestion might appear similar to the requirement of providing two independent sets of inferential statistics on the same basic effect, at the core of most replicability efforts (Lindsay, 2017). However, the estimation stage is not about confirming the “truth” of an exploratory observation, already controlled for by the nominal rate of false positives (e.g., $\alpha = 5\%$) or equivalently corrected when multiple tests are performed. The rationale for providing a second, independent set of observations is to increase the precision of effect size determination for a finding deemed interesting enough to justify additional and substantial measurement efforts, and to consider whether the magnitude of that more precise estimate is biologically relevant. In this context, the estimation phase would benefit from registration, since it is important to document the precise replication of the experimental protocol and analytical procedures of the exploratory stage. This is especially important if first-stage estimation experiments are not published. This procedure should lead researchers to expect, rather than criticize, smaller effect sizes in the estimation stage. Those smaller effect sizes, combined with low power in the exploratory stage, will result in many estimation studies failing to confirm the rejection of the null hypothesis. However, rather than jumping to the conclusion that the inferences of the estimation stage were “false” (Ioannidis, 2005), this two-step procedure might shift the emphasis toward precisely estimating the magnitude and direction of an effect (“how much”) and away from a dichotomous (“Does the effect exist or not?”) question (Calin-Jageman and Cumming, 2019). Put differently, this invites researchers to evaluate the biological plausibility of more precisely estimated effects, rather than use an inferential threshold as a license to suspend critical judgment (Gigerenzer, 2018).

The suggested exploration-then-estimation procedure is functionally equivalent to practices already adopted by some subfields of neuroscience. For instance, in cognitive neuroscience, it is customary to separate the estimation phase of model fitting from the validation phase of the model parameters. That approach is valid as long as the validation phase operates on independent data and does not introduce new parameters. While many of these practices typically rely on large sample sizes, some areas of neuroscience make statistical inferences on individual subjects, implementing a sort of exploration-then-estimation procedure across successive subjects (e.g., patients or nonhuman animal models in electrophysiology; machine-learning explorations of fMRI data; psychophysics and human brain lesion studies). These small-

N approaches focus their statistical power on individual-level characterization of an effect; a finding is deemed present when all or a majority of a small pool of subjects show an effect, usually based on a large sample of trial-level observations (Smith and Little, 2018). It should be acknowledged that this approach only allows for statements that pertain to the existence and magnitude of effects in those subjects, rather than in the populations those subjects are drawn from. Many of the most robust findings in psychophysics have come from a small-N approach (Smith and Little, 2018), and it could be preferred ethically when animal welfare or vulnerable individuals are involved.

In addition to screening submissions for the rigor of their statistical procedures, we believe it is also important to steer the community through positive examples. *JNeurosci* welcomes contributions that provide a definite statement on a research question by commenting on biological plausibility and using rigorous statistical procedures, such as those discussed in this editorial.

We invite you to contribute to this discussion by emailing *JNeurosci* at JN_EiC@sfn.org or tweeting to @marinap63.

The Editorial Board of *The Journal of Neuroscience*
<https://doi.org/10.1523/JNEUROSCI.0866-20.2020>

References

- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376.
- Calin-Jageman RJ, Cumming G (2019) Estimation for better inference in neuroscience. *eNeuro* 6:ENEURO.0205-19.2019.
- Gigerenzer G (2018) Statistical rituals: the replication delusion and how we got there. *Adv Methods Pract Psychol Sci* 1:198–218.
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:e124.
- Lindsay DS (2017) Preregistered direct replications in psychological science. *Psychol Sci* 28:1191–1192.
- Maxwell SE, Kelley K, Rausch JR (2008) Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 59:537–563.
- Picciotto M (2018) Analytical transparency and reproducibility in human neuroimaging studies. *J Neurosci* 38:3375–3376.
- Smith PL, Little DR (2018) Small is beautiful: in defense of the small-N design. *Psychon Bull Rev* 25:2083–2101.
- Wilson BM, Harris CR, Wixted JT (2020) Science is not a signal detection problem. *Proc Natl Acad Sci U S A* 117:5559–5567.