# Scientific articles and peer review

Natalia Zaretskaya

SS 2025

# Table of contents

# Overview

In this first block of the course we will learn all about scientific articles

# Assessing a study sample

## Ideal sample

What are the characteristics of an ideal sample

## Sample parameters

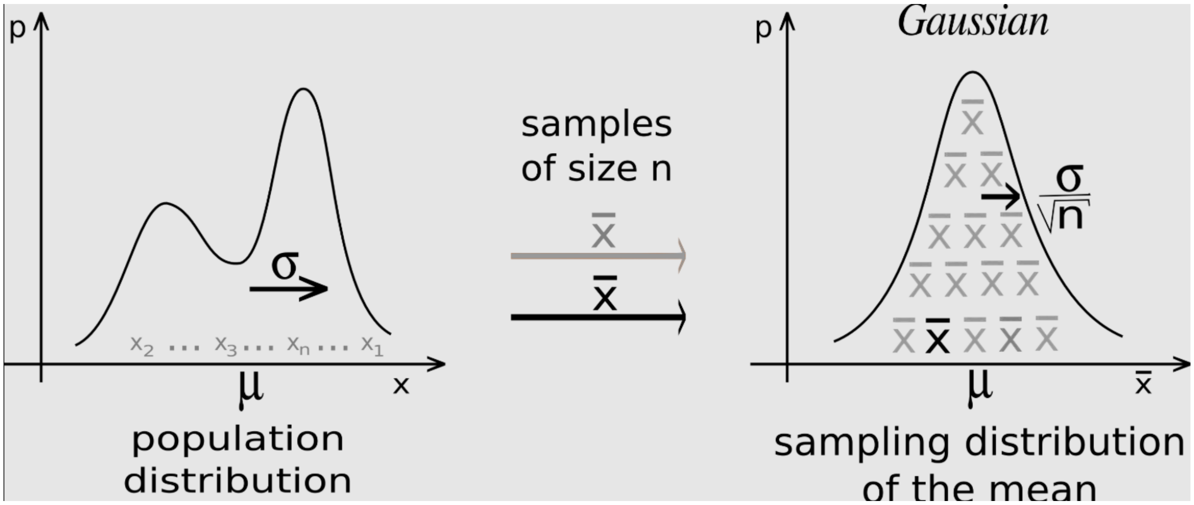
- Size
- Composition

## Sample composition

- The need for stratified sampling depends on the research question
- In basic research, it is common (and valid!) to use healthy young adults
- More problematic: psychology students

# Sample size

## Central limit theorem


p
σ
$x_2$ ... $x_3$ ... $x_n$ ... $x_1$
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
Gaussian
p
$\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
μ
$\bar{x}$
sampling distribution of the mean

[1]

The central limit theorem makes two claims.

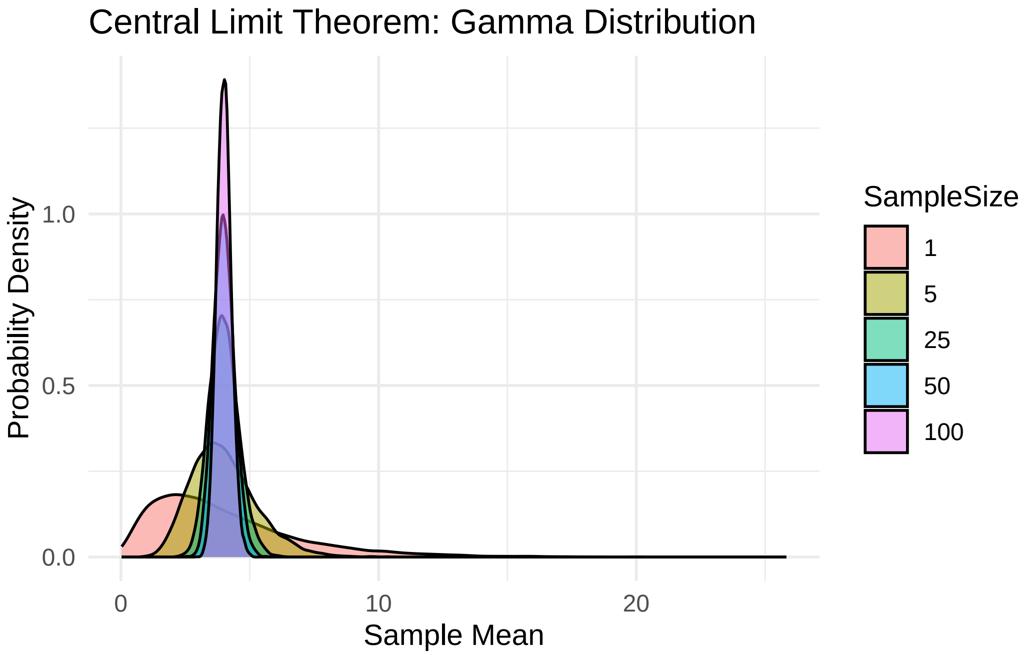
Claim 1: - If the population distribution of some parameter is not normal, - And your draw a random sample of size N from this distribution for your study, - If you were to repeat the same same study many times, the distribution of sample *means* will approach the normal distribution. This means that you can apply parametric statistical tests even for non-normally distributed data.

Claim 2: The larger your sample size, the better its mean represents the population mean (the smaller are the confidence intervals)

[1] CC BY-SA 4.0 Mathieu ROUAUD

## Illustration

```r
# Load necessary library
library(ggplot2)

# Function to simulate sample means and plot distributions
simulate_clt <- function(n_values, num_samples = 10000, shape = 2, scale = 2) {
  sample_means <- list()

  for (n in n_values) {
    means <- replicate(num_samples, mean(rgamma(n, shape = shape, scale = scale)))
    sample_means[[as.character(n)]] <- means
  }

  # Convert to data frame for ggplot
  data <- do.call(rbind, lapply(names(sample_means), function(n) {
    data.frame(SampleMean = sample_means[[n]], SampleSize = as.factor(n))
  }))

  # Plot the distributions of sample means
  ggplot(data, aes(x = SampleMean, fill = SampleSize)) +
    geom_density(alpha = 0.5) +
    labs(title = "Central Limit Theorem: Gamma Distribution",
         x = "Sample Mean", y = "Probability Density") +
    theme_minimal()
}

# Define sample sizes to test
sample_sizes <- c(1, 5, 25, 50, 100)

# Run the simulation
simulate_clt(sample_sizes)
```



## Low sample sizes

| Sample size | Type I error / / false positive rate | Type II error / / false negative rate |
|---|---|---|
| small | pre-set to 0.05 | high |
| large | pre-set to 0.05 | low |

Power: probability to detect a significant effect in the data if there really is one in the population

$$Power = (1 - \boldsymbol{\beta})$$

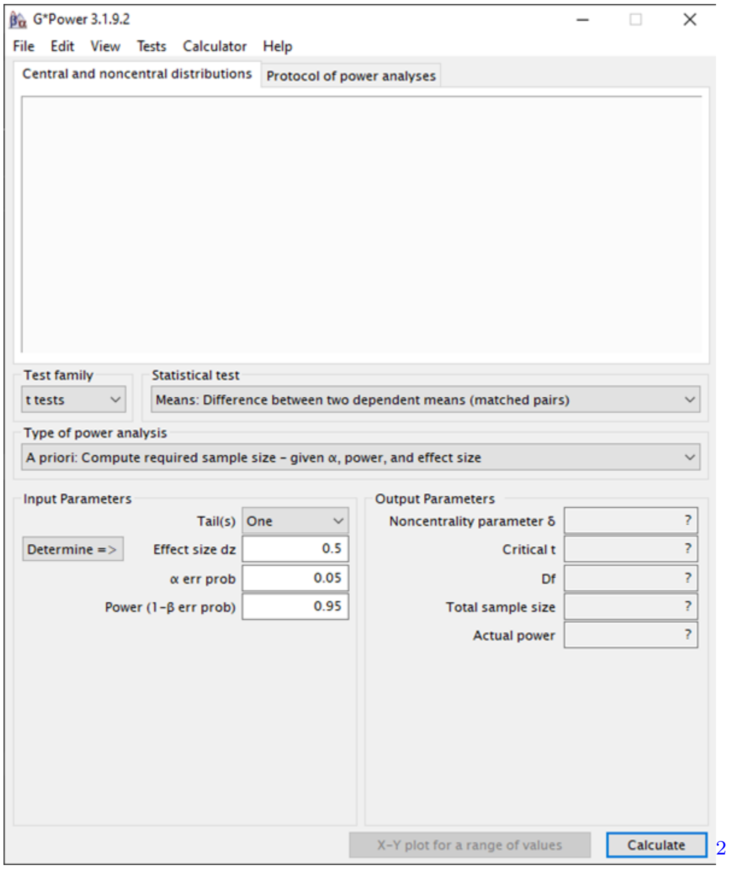
Low sample size does not mean you can't trust the reported effect. The false positive rate (aka alpha, aka type I error) is always pre-set to a specific value (usually p=0.05).

Low sample size reducers the power (probability of detecting an effect in the population), aka (1-beta), aka 1-type II error, aka false negative.

## Inference with low sample size

- Normality assumption has to be tested
- Non-parametric tests should be applied if the test is not passed

## *a priori* power analysis


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

| | |
|---|---|
| Tail(s) | One |
| Effect size dz | 0.5 |
| α err prob | 0.05 |
| Power (1-β err prob) | 0.95 |

Output Parameters

| | |
|---|---|
| Noncentrality parameter δ | ? |
| Critical t | ? |
| Df | ? |
| Total sample size | ? |
| Actual power | ? |

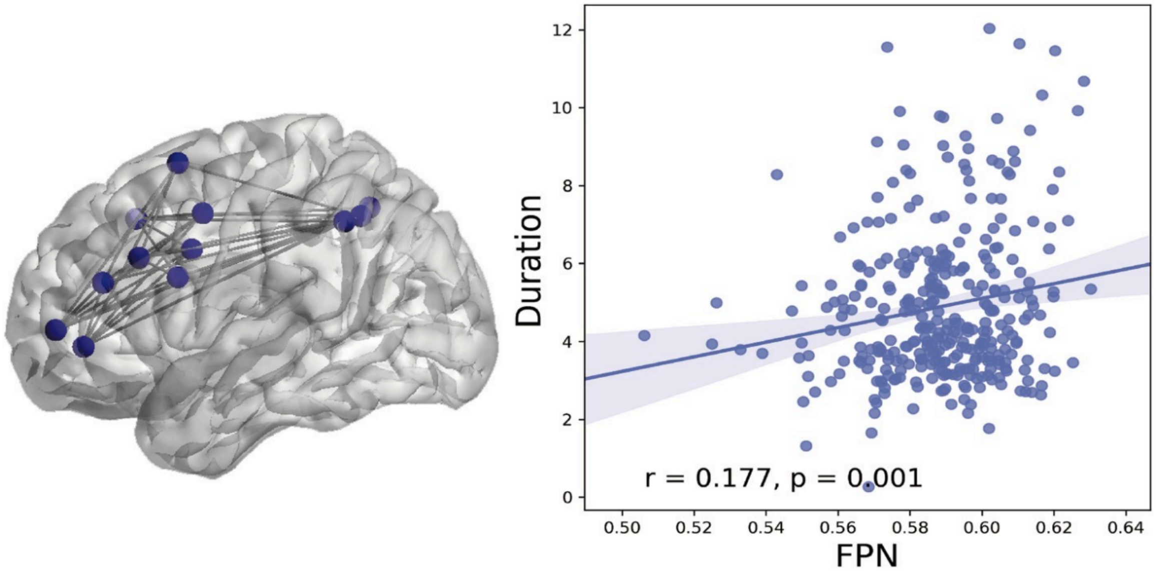
[2]

Information that goes into the formula:

- Expected effect size
- Statistical test type
- Desired p-value to be exceeded (aka type I error, aka )
- *Sample size*

[2] Faul et al. (2007); also possible in SPSS and R

## Disadvantages of large samples[3]



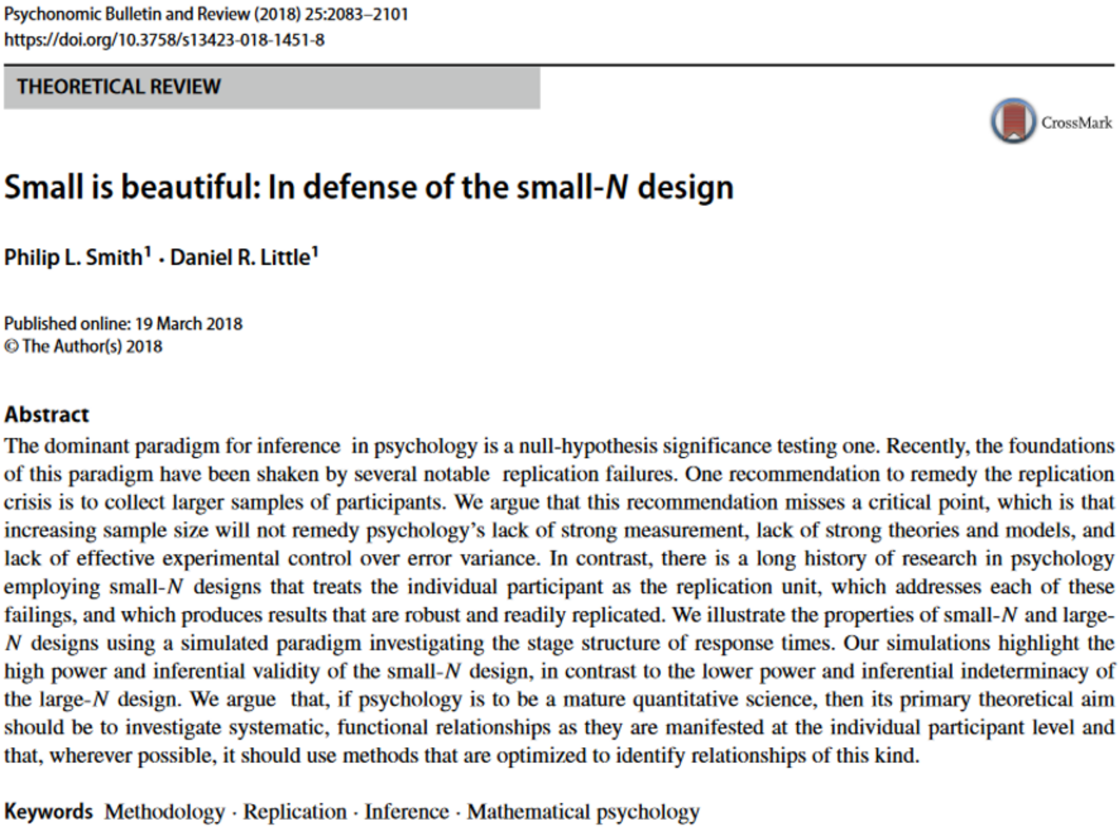
Tiny correlation coefficients easily become significant with large samples. It is, however, a question, whether small effects have any biological relevance.

## Take-home messages

- There is no magic number that distinguishes insufficient form sufficient sample size, it is a continuum
- Normality tests need to be performed for studies with low sample size
- Choice of sample size should to be (ideally) substantiated by an *a priori* power analysis
- Post-hoc power analysis is less relevant, if the effect size observed in the data is being used

[3] Mao et al. (2020)

## Reading recommendation[4]



Psychonomic Bulletin and Review (2018) 25:2083–2101
https://doi.org/10.3758/s13423-018-1451-8

THEORETICAL REVIEW

### Small is beautiful: In defense of the small-*N* design

Philip L. Smith[1] · Daniel R. Little[1]

Published online: 19 March 2018

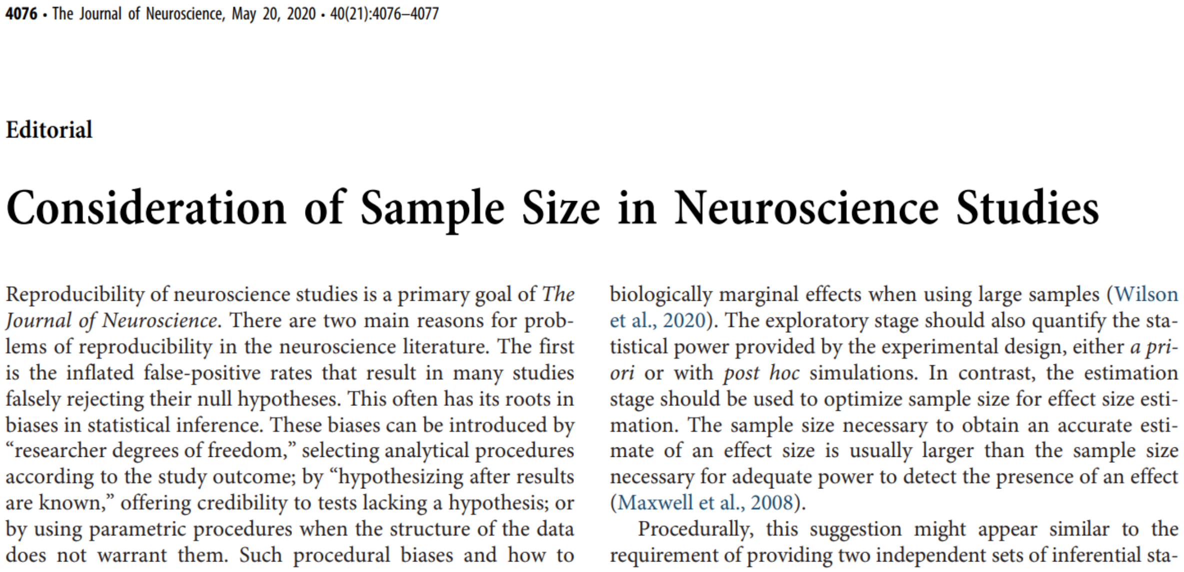
**Abstract**
The dominant paradigm for inference in psychology is a null-hypothesis significance testing one. Recently, the foundations of this paradigm have been shaken by several notable replication failures. One recommendation to remedy the replication crisis is to collect larger samples of participants. We argue that this recommendation misses a critical point, which is that increasing sample size will not remedy psychology's lack of strong measurement, lack of strong theories and models, and lack of effective experimental control over error variance. In contrast, there is a long history of research in psychology employing small-*N* designs that treats the individual participant as the replication unit, which addresses each of these failings, and which produces results that are robust and readily replicated. We illustrate the properties of small-*N* and large-*N* designs using a simulated paradigm investigating the stage structure of response times. Our simulations highlight the high power and inferential validity of the small-*N* design, in contrast to the lower power and inferential indeterminacy of the large-*N* design. We argue that, if psychology is to be a mature quantitative science, then its primary theoretical aim should be to investigate systematic, functional relationships as they are manifested at the individual participant level and that, wherever possible, it should use methods that are optimized to identify relationships of this kind.

**Keywords** Methodology · Replication · Inference · Mathematical psychology

These are some considerations from the field of basic vision science, were small samples were not unusial (this changes though). The logic of performing statistical inference on an individual-subject level, treating each subject as a replication unit, is valid in other fields of neuroscience as well.

[4] Smith and Little (2018)

## Reading recommendation[5]



**4076** • The Journal of Neuroscience, May 20, 2020 • 40(21):4076–4077

**Editorial**

# Consideration of Sample Size in Neuroscience Studies

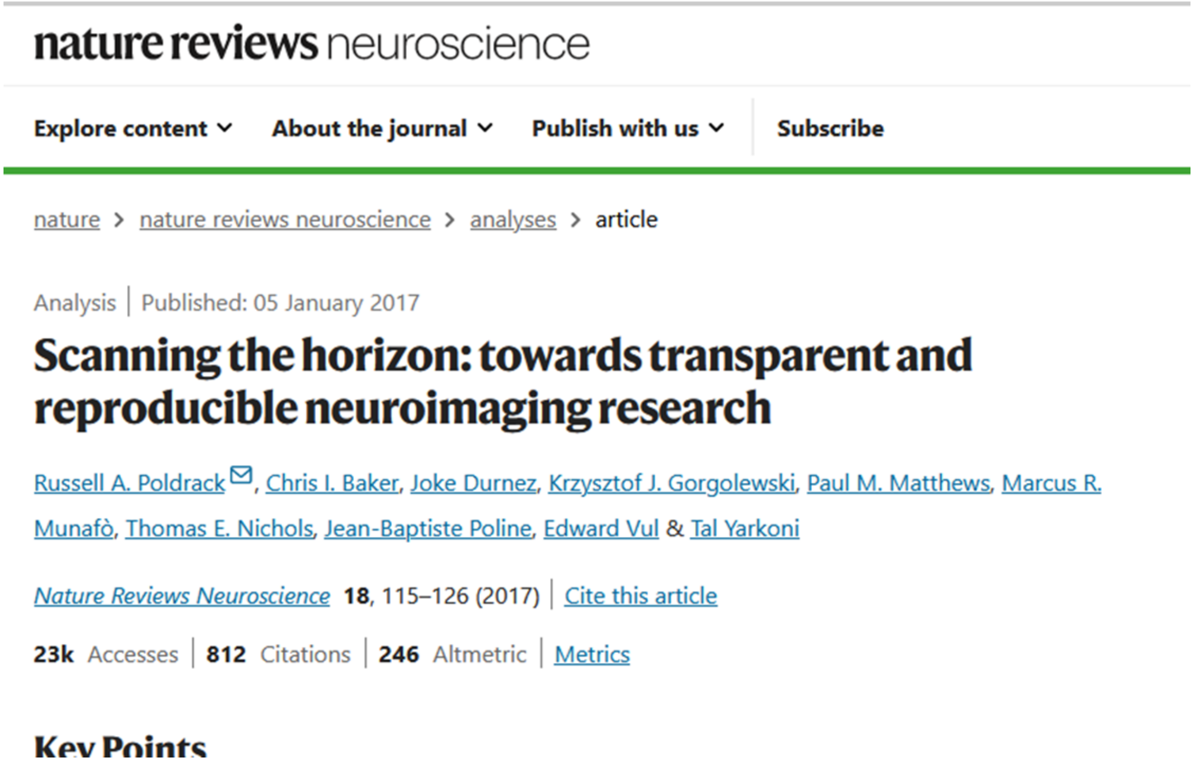
Reproducibility of neuroscience studies is a primary goal of *The Journal of Neuroscience*. There are two main reasons for problems of reproducibility in the neuroscience literature. The first is the inflated false-positive rates that result in many studies falsely rejecting their null hypotheses. This 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 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 sta-

These are some considerations for neuroscience studies, which often involve animals. One would want to minimize the number of animals used for research, while still being able to make conclusions about the presence of an effect.

[5] "Consideration of Sample Size in Neuroscience Studies" (2020)

## Reading recommendation[6]


nature reviews neuroscience
Explore content
About the journal
Publish with us
Subscribe
nature > nature reviews neuroscience > analyses > article

Analysis | Published: 05 January 2017

**Scanning the horizon: towards transparent and reproducible neuroimaging research**

Russell A. Poldrack, Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul & Tal Yarkoni

*Nature Reviews Neuroscience* **18**, 115–126 (2017) | Cite this article

**23k** Accesses | **812** Citations | **246** Altmetric | Metrics

**Key Points**

These are some considerations that are specific to neuroimaging studies

## References

"Consideration of Sample Size in Neuroscience Studies." 2020. *The Journal of Neuroscience* 40 (21): 4076–77. https://doi.org/10.1523/jneurosci.0866-20.2020.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. https://doi.org/10.3758/bf03193146.

Mao, Yu, Ryota Kanai, Cody Ding, Taiyong Bi, and Jiang Qiu. 2020. "Temporal Variability of Brain Networks Predicts Individual Differences in Bistable Perception." *Neuropsychologia* 142 (May): 107426. https://doi.org/10.1016/j.neuropsychologia.2020.107426.

Poldrack, Russell A., Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward

[6] Poldrack et al. (2017)

Vul, and Tal Yarkoni. 2017. "Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research." *Nature Reviews Neuroscience* 18 (2): 115–26. https://doi.org/10.1038/nrn.2016.167.

Smith, Philip L., and Daniel R. Little. 2018. "Small Is Beautiful: In Defense of the Small-N Design." *Psychonomic Bulletin & Review* 25 (6): 2083–2101. https://doi.org/10.3758/s13423-018-1451-8.