Sample size estimation for task-related functional MRI studies using Bayesian updating

Eduard T. Klapwijk Joran Jongerling Herbert Hoijtink
Eveline A. Crone

2024-06-13

Task-related functional MRI (fMRI) studies need to be properly powered with an adequate sample size to reliably detect effects of interest. But for most fMRI studies, it is not straightforward to determine a proper sample size using power calculations based on published effect sizes. Here, we present an alternative approach of sample size estimation with empirical Bayesian updating. First, this method provides an estimate of the required sample size using existing data from a similar task and similar region of interest. Using this estimate researchers can plan their research project, and report empirically determined sample size estimations in their research proposal or pre-registration. Second, researchers can expand the sample size estimations with new data. We illustrate this approach using four existing fMRI data sets where Cohen's d is the effect size of interest for the hemodynamic response in the task condition of interest versus a control condition, and where a Pearson correlation between task effect and age is the covariate of interest. We show that sample sizes to reliably detect effects differ between various tasks and regions of interest. We provide an R package to allow researchers to use Bayesian updating with other task-related fMRI studies.

1 Introduction

Since the emergence of functional magnetic resonance imaging (fMRI), these techniques have provided unprecedented opportunities to study functional brain development during childhood and adolescence by scanning children from the age of four years onward. There is great progression in the assessment of neural functional growth using cross-sectional and longitudinal assessments of cognitive, social and affective processes across the full range of childhood to adulthood. Despite the advancements in the ability to study the developing brain in vivo using fMRI, recent years have seen an increased concern about the replicability of scientific findings in general and particularly in psychology and cognitive neuroscience (Bishop 2019; E. T. Klapwijk et al. 2021; Munafò et al. 2017; Poldrack et al. 2017; see Marek et al. 2022 for a similar concern for resting-state functional connectivity and structural MRI data sets).

Low statistical power due to small sample sizes is arguably one of the main reasons for lower than desired replicability of study results. Statistical power is the probability that a study will reject the null hypothesis when it is false, that is, when there is a non-zero effect (e.g., Cohen's d or Pearson's correlation) in the population of interest. Power is determined by the size of the effect, the alpha level chosen, and the size of the sample (Cohen 1992). The smaller each of effect size, alpha level, and sample size are, the lower the power. Since many psychological phenomena consist of relatively subtle, small effects (Funder and Ozer 2019; Gignac and Szodorai 2016; Yarkoni 2009), the main source for increasing power, and over which researchers have a reasonable degree of control, is the sample size (given limited flexibility of the alpha level). Despite the need for well-powered studies to reliably detect effects of interest, empirical reports have shown that most cognitive neuroscience and fMRI studies are underpowered due to small sample sizes (Button et al. 2013; Maxwell 2004; Nord et al. 2017; Poldrack et al. 2017; Szucs and Ioannidis 2017; Turner et al. 2018). With developmental populations, sufficiently large sample sizes might be even harder to establish because of the challenges in recruitment and testing of young participants (Achterberg and Meulen 2019; E. T. Klapwijk et al. 2021).

Recently, well-coordinated efforts and funding have led to several large-scale projects that collect developmental task-related fMRI data with larger sample sizes. These projects have the potential to resolve the problem of power with sample sizes in the thousands, such as the IMAGEN study ($N \approx 2,000$) (Schumann et al. 2010), the Philadelphia Neurodevelopmental Cohort ($N \approx 1,000$) (Satterthwaite et al. 2016), the Human Connectome Project in Development (HCP-D; $N \approx 1,300$) (Somerville et al. 2018), and the Adolescent Brain Cognitive Development (ABCD) study ($N \approx 11,000$) (Casey et al. 2018). The leap forward made with this wealth of high-powered, mostly publicly available data can hardly be overstated, given that they provide an important open research tool for researchers across the globe.

However, most fMRI studies are carried out by individual research groups in much smaller samples, which have more opportunities to pursue new scientific questions, for example using novel paradigms. It is also vital that the findings of such studies are replicable and meaningful, meaning that these studies should be properly powered, also without sample sizes in the range of large multi-lab studies. The main issue with power analysis is that the effect size in the population of interest is unknown. One option is to use effect sizes reported in the literature of the research area at hand. But these effect sizes are often inflated due to publication bias (Gelman and Carlin 2014; Ioannidis 2005; Open Science Collaboration 2015; Wicherts et al. 2016; Yarkoni 2009). Therefore, calculating power based on published effect sizes usually underestimates the sample size needed to reliably detect an effect.

This paper will present a novel method to determine the required sample size for fMRI studies based on existing (i.e., already collected) data using Bayesian updating (Rouder 2014). Specifically, the approach will determine the proportion of the already collected data that is needed to get a desired credible interval (the Bayesian counterpart of the confidence interval). This will provide an estimate of the percentage of cases for which the credible interval is expected to be in the desired range (e.g., the interval should *not* contain the value 0 for the parameter of interest). This in turn gives us an estimate of the sample size needed for a certain level of power. The current paper will provide examples (including an R package) for two effect sizes: Cohen's d and Pearson's correlation. The sample size determined using existing data is valuable when designing a new research project and when justifying sample sizes in a pre-registration or in proposals send to a (medical) ethical committee.

We will illustrate sample size determination using existing data sets and tasks that are currently widely used in the developmental fMRI literature, specifically cognitive control, reward activity, and social-cognitive processing (see Table 1 for an overview) based on existing data from our own lab. It will be determined how large the sample size should be to detect Cohen's d, such that 95% does not contain the value zero for a specific condition effect (e.g., brain activity during feedback processing versus rule application). Most prior developmental fMRI studies addressed the question whether an effect linearly increases or decreases with age (Crone and Steinbeis 2017). We therefore also determine the sample size needed to detect a Pearson correlation of an effect with linear age that is larger than zero. In the next sections, we will first introduce Bayesian updating, the highest density credible interval, and sample size determination. Next, we will provide examples using existing data from four fMRI studies (Braams et al. 2014; Peters and Crone 2017; Spaans et al. 2023; Cruijsen et al. 2023), and illustrate sample size determination using these examples.

1.1 Bayesian updating

Bayesian updating can be used to determine the sample size required to estimate Cohen's d or Pearson's correlation with a certain precision. Precision is presented in the form of a 95% highest density credible interval (HDCI), that is, the narrowest possible interval that is believed to contain the true value of the parameter of interest with a probability of .95[1]. The narrower this interval, the higher the precision with which the parameter is estimated.

Bayesian updating as implemented here relies on the assumption that a priori, that is, before collecting the data, each possible value of the parameter of interest is equally likely. This has two implications. First, the HDCI is completely determined by the data and not by prior information. Secondly, the numerical values of the estimate and endpoints of the HDCI are equal to the estimate and endpoints of the classical confidence interval.

Bayesian updating can be used to determine the smallest sample size for which the resulting HDCI does not contain the value zero. Bayesian updating consists of four steps:

1. Determine the maximum achievable size of the sample.

[1] In statistical terms: the HDCI contains 95% of the marginal posterior density of the parameter which is a function of the information in the data and prior knowledge with respect to the parameter.

2. Collect the data for an initial number of participants and then compute the estimate and HDCI. The actual number chosen is irrelevant, but with 20 participants the estimate and HDCI will usually give a first impression of the size of Cohen's d or Pearson's correlation that is not totally determined by sample fluctuations and/or outliers.

3. Add several participants to the sample (updating) and recompute the estimate and HDCI.

4. If the HDCI still contains the value zero and the maximum achievable sample size has not been obtained, return to the previous step. Otherwise, the updating is finished and estimates and corresponding HDCI's are reported.

1.1.1 The highest density credible interval (HDCI)

It is important to highlight that the HDCI is not a confidence interval. If many data sets are sampled from the population of interest and each data set is used to compute the 95% confidence interval for the parameter of interest, then it holds that 95% of the intervals contain the true value. However, in contrast to HDCIs, confidence intervals cannot be updated because their coverage level will become smaller than 95%. This will be illustrated using a simple example.

Imagine many researchers want to show that Cohen's d is unequal to zero. Each of them samples a data set with $N = 20$ for each group from the population of interest (in which Cohen's d happens to equal zero). About 5% of the 95% confidence intervals do not contain the value 0. These researchers will not continue their efforts; they have "shown" that "their" effect is not zero. At this stage the Type I error rate is .05. However, the 95% remaining researchers increase their power by updating their data with another 10 persons per group and recompute their 95% confidence intervals of which about 2.8% (number determined using simulation) does not contain the value 0. Therefore, the Type I error rate is increased to 5% + 2.8% = 7.8%, that is, the updating rendered 92.2% confidence intervals and not 95% confidence intervals.

A HDCI should therefore not be interpreted in the context of many hypothetical data sets sampled from the population of interest. The HDCI is computed using the observed data at hand and is the shortest interval that is believed to contain the true value of Cohen's d with a probability of .95. When updating, the size of the data at hand becomes larger, the information with respect to Cohen's d increases, and therefore the width of the HDCI becomes smaller. The HDCI summarizes the evidence in the data at hand with respect to the size of Cohen's d, which is different from a confidence interval which aims to control error rates under (hypothetical) repeated sampling from the same population.

1.2 Sample Size Determination

The procedure elaborated in this paper is Bayesian (because HDCIs are used) empirical (because existing data are used) sample size determination. We use an existing dataset of a certain size, and for all sample sizes between a minimum (e.g., $N = 20$) and maximum sample size of the sample at hand we compute the HDCI. To account for the arbitrary order of the participant in a data set, this is done for a set number of permutations, e.g., 1000, of the participants in the data set. For each sample size, we then calculate the average HDCI of all permutations. Considering both the average HDCI and the HDCI's resulting from the different permutations, will provide an estimate sample size needed to obtain a 95% HDCI . The 95% HDCI for Cohen's d was established through non-parametric bootstrapping (Efron and Tibshirani 1994). Specifically, we resampled the current subset of data 100 times and calculated Cohen's d for the difference between variable x and variable y each time. We then calculated the mean and standard deviations of Cohen's d across the 100 resampled values. The 95% CI was determined by subtracting 1.96 times the standard deviation of the Cohen's d values from the mean Cohen's d value, and by adding 1.96 times the standard deviation of the Cohen's d values to the mean Cohen's d value.

To illustrate the current method, we can take a preview at Figure 1 C in the Results section. The dataset used to construct Figure 1 C HDCI's for Cohen's d of 149 participants from the self-evaluation versus control contrast in the medial prefrontal cortex during a self-evaluation task. In Figure 1 C the first 20 participants in each of 1000 permuted data sets are used to compute the HDCI for Cohen's d. Ten of these intervals are displayed in blue. As can be seen, of the ten (of 1000) HDCIs displayed, four include the value zero. Consequently, fluctuations in the order in which participants are sampled determine whether the HDCI contains the value zero. This is summarized in Figure 2 C which shows that the probability that one of the 1000 HDCIs does not contain the value zero is about 0.6 for $N = 20$. Figure 1 C (the first interval displayed in purple) show that the average of the 1000 HDCI's do not contain the value zero. Therefore, although the average interval does not contain the value zero, it is learned from the permutations that when using 20 participants it is still rather likely that the resulting interval will include the value zero. This becomes much less likely for the samples of 46 and 72 participants. With these sample sizes neither the HDCIs nor the average interval contain the value zero (see Figure 1 C). Additionally, the probability that an interval does not contain the value zero equals 100% for 72 participants (see Figure 2 C). Therefore, a sample size of 72 and higher will usually render intervals that do not contain the value zero.

An issue that sample size determination has in common with power analysis, is that the effect size in the existing data set will very likely differ from the effect size in the population that will be addressed in the study being designed. However, sample size determination does not suffer from the fact that effect sizes reported in the literature tend to be inflated (like power analysis) because effect sizes are straightforwardly computed from actual data. The determined sample size is therefore an estimate of the required sample size. This estimate is valuable because it allows researchers to plan their research project, can be reported in their research proposal or pre-registration, and it may lead to the conclusion that the sample size needed cannot be achieved with the resources that you have at your disposal.

In the next section Bayesian empirical sample size determination will be exemplified and discussed using different fMRI tasks from the BrainTime (Braams et al. 2014; Peters and Crone 2017) and Self-Concept (Spaans et al. 2023; Cruijsen et al. 2023) data sets.

1.3 *neuroUp* R package

The code for Bayesian updating in this manuscript is implemented in the R language for statistical computing (R Core Team, 2022). We currently provide the *neuroUp* package at the following location: https://github.com/eduardklap/neuroUp (E. Klapwijk, Hoijtink, and Jongerling 2024). This package is built on several packages from the tidyverse (Wickham et al. 2019), most notably *ggplot2*. The package can be installed using the R commands: `library(devtools)` and `install_github("eduardklap/neuroUp")`. All data used in the current manuscript is also available within the *neuroUp* package, in this way all figures in the manuscript can be reproduced using the R package. For a more elaborate introduction to *neuroUp* and its functions, please refer to https://eduardklap.github.io/neuroUp/articles/neuroUp.html. To cite package neuroUp in publications use: Klapwijk, E., Hoijtink, H., & Jongerling, J. (2024). neuroUp: Plan sample size for fMRI regions of interest research using Bayesian updating. https://doi.org/10.5281/zenodo.11526169

2 Method & Materials

3 Results

3.1 Task effects using Cohen's d

The average HDCI was used to determine the optimal sample size for the four fMRI tasks used. For each task and brain region of interest, we estimated the sample size required to obtain an average HDCI for Cohen's d not containing the value zero (Figure 1 and 2; Table 2). Additionally, we also established whether it was also the case that a majority of the 1000 underlying HDCIs did not contain the value zero.

In Figure 1 A, an estimate of the required sample size for feedback learning processing in the DLPFC (middle frontal gyrus) is presented. Five groups of HDCI's are presented for sample sizes of 20, 70, 120, 170, and 271 persons, in which 271 is the total group sample size of the existing data set. As can be seen, already with a sample of 20 participants neither of the 10 permuted HDCIs displayed nor the average of the 1000 HDCI's contain the value zero. In Figure 2A it can be seen that the proportion of HDCI's not containing the value zero is equal to 1.0. This is due to the huge effect size of Cohen's d equal to about 1.9 (see Table

2). Therefore, already with a sample size of 20 participants, an effect bigger than zero will be detected for this task and brain region.

For the task effect of gambling (winning for self > losing for self contrast) in the NAcc, we can see the results in Figure 1 B. Here, with a sample of 20 participants the average of the 1000 HDCIs does not contain the value zero but some of the 10 HDCI's displayed do. The proportion of HDCI's not containing the value zero is bigger than 0.8 for 20 participants, but with 60 or more participants none of the 1000 HDCI's contain zero anymore (Figure 2B). This is related to the large effect size of Cohen's d equal to about 0.7 (see Table 2). Thus, for this task and brain region, already with a sample size of 20 participants chances are high that an effect bigger than zero will be detected. Using sample sizes of 60 or more participants from this existing data set increases the chances of detection to 100%.

In Figure 1 C, results are plotted for the self-evaluation task in the mPFC. We see that at a sample size of 20 some of the 10 HDCI's displayed still contain the value 0, but not the average of the 1000 HDCI's. The proportion of HDCI's not containing the value 0 is also below 1.0, around 0.5, for this task at $N = 20$ (Figure 2C). At the next step that is plotted for this task ($N = 46$), we see that the proportion of HDCI's not containing the value 0 is only a little below 1.0. Thus, for this task and brain region, at a sample size at 46 participants chances are high that an effect bigger than zero will be detected.

The results of gaining for self in the NAcc are plotted in Figure 1 D. With a sample of 20 participants the average of the 1000 HDCI's does not contain the value zero but some of the 10 HDCI's displayed do. The proportion of HDCI's not containing the value zero is around 0.75 (Figure 2D), but at $N = 47$ the proportion of HDCI's not containing the value 0 is almost 1.0. Therefore, using 20 participants it is still likely that the resulting interval will include the value zero. This becomes much less likely for the samples of 47 and 74 persons. With these sample sizes neither the HDCIs nor the average interval contain the value zero. Additionally, the probability that an interval does not contain the value zero almost reaches 100%. Therefore, for this task and brain region, a sample size of 47 and higher will usually render intervals that do not contain the value zero.

Table 1: **Table 2.** Mean estimates (with credible interval in brackets) of Cohen's d for five different sample sizes (starting with N=20, then 1/5th parts of the total dataset) of the 1000 HDCI's. DLPFC = dorsolateral prefrontal cortex; mPFC = medial prefrontal cortex; NAcc = nucleus accumbens.

task	**brain region**	**N = 20**	**N = 2/5**	**N = 3/5**	**N = 4/5**	**N = total**
Feedback	DLPFC	**2.03** (1.29, 2.76)	**1.89** (1.54, 2.25), $N = 70$	**1.88** (1.62, 2.15), $N = 120$	**1.87** (1.65, 2.1), $N = 170$	**1.87** (1.69, 2.04), $N = 271$
Gambling	NAcc	**0.8** (0.25, 1.34)	**0.74** (0.45, 1.03), $N = 60$	**0.73** (0.51, 0.96), $N = 100$	**0.73** (0.54, 0.92), $N = 140$	**0.73** (0.58, 0.88), $N = 221$

Self-evaluations	mPFC	**0.5** (0.02, 0.98)	**0.47** (0.17, 0.77), N = 46	**0.46** (0.22, 0.7), N = 72	**0.46** (0.26, 0.66), N = 98	**0.46** (0.29, 0.62), N = 149
Gaining self	NAcc	**0.64** (0.14, 1.14)	**0.59** (0.27, 0.91), N = 47	**0.58** (0.32, 0.84), N = 74	**0.58** (0.36, 0.8), N = 101	**0.57** (0.39, 0.75), N = 156

3.2 Data & Methods

3.3 Conclusion

References

Achterberg, Michelle, and Mara van der Meulen. 2019. "Genetic and Environmental Influences on MRI Scan Quantity and Quality." *Developmental Cognitive Neuroscience* 38 (August): 100667. https://doi.org/10.1016/j.dcn.2019.100667.

Bishop, Dorothy. 2019. "Rein in the Four Horsemen of Irreproducibility." *Nature* 568 (7753): 435–35. https://doi.org/10.1038/d41586-019-01307-2.

Braams, Barbara R., Sabine Peters, Jiska S. Peper, Berna Güroğlu, and Eveline A. Crone. 2014. "Gambling for Self, Friends, and Antagonists: Differential Contributions of Affective and Social Brain Regions on Adolescent Reward Processing." *NeuroImage* 100 (October): 281–89. https://doi.org/10.1016/j.neuroimage.2014.06.020.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Casey, B. J., Tariq Cannonier, May I. Conley, Alexandra O. Cohen, Deanna M. Barch, Mary M. Heitzeg, Mary E. Soules, et al. 2018. "The Adolescent Brain Cognitive Development (ABCD) Study: Imaging Acquisition Across 21 Sites." *Developmental Cognitive Neuroscience*, The adolescent brain cognitive development (ABCD) consortium: Rationale, aims, and assessment strategy, 32 (August): 43–54. https://doi.org/10.1016/j.dcn.2018.03.001.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 155–59. https://doi.org/10.1037/0033-2909.112.1.155.

Crone, Eveline A., and Nikolaus Steinbeis. 2017. "Neural Perspectives on Cognitive Control Development during Childhood and Adolescence." *Trends in Cognitive Sciences* 21 (3): 205–15. https://doi.org/10.1016/j.tics.2017.01.003.

Cruijsen, Renske van der, Neeltje E Blankenstein, Jochem P Spaans, Sabine Peters, and Eveline A Crone. 2023. "Longitudinal Self-Concept Development in Adolescence." *Social Cognitive and Affective Neuroscience* 18 (1): nsac062. https://doi.org/10.1093/scan/nsac062.

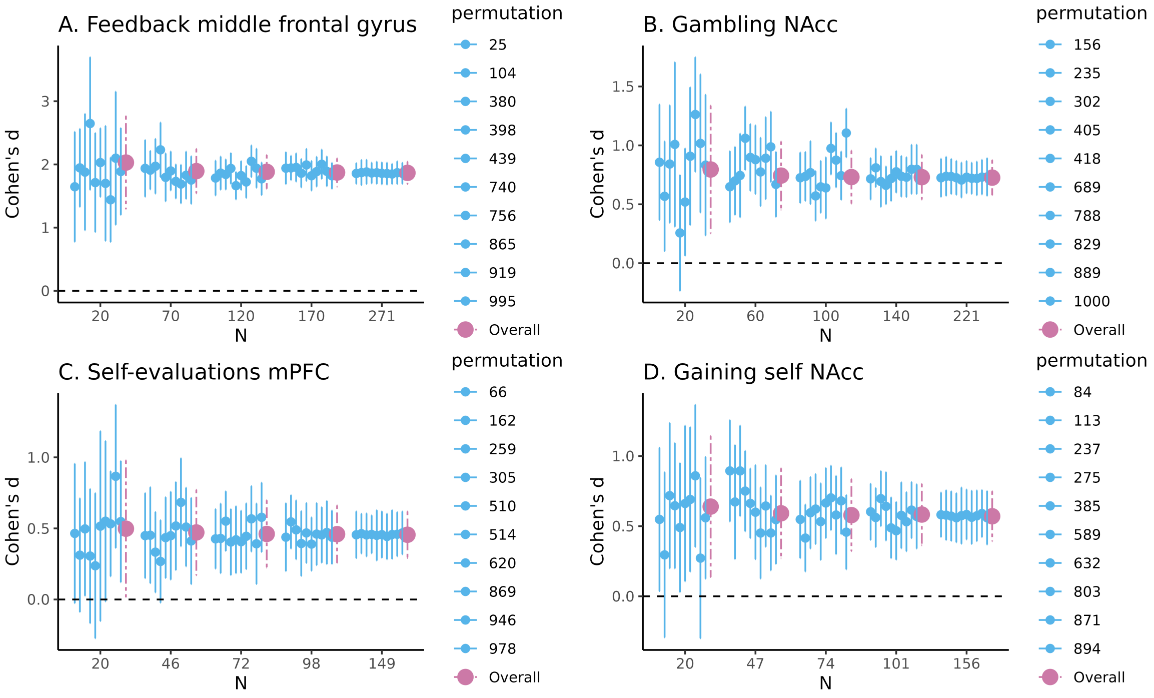

Figure 1: Estimates of task effects for five different sample sizes (starting with $N = 20$, then 1/5th parts of the total dataset). For each sample size 10 out of the 1000 HDCIs computed are displayed (in light blue). The average estimate with credible interval summarizing the 1000 HDCIs for each sample size are plotted in reddish purple. mPFC = medial prefrontal cortex; NAcc = nucleus accumbens.

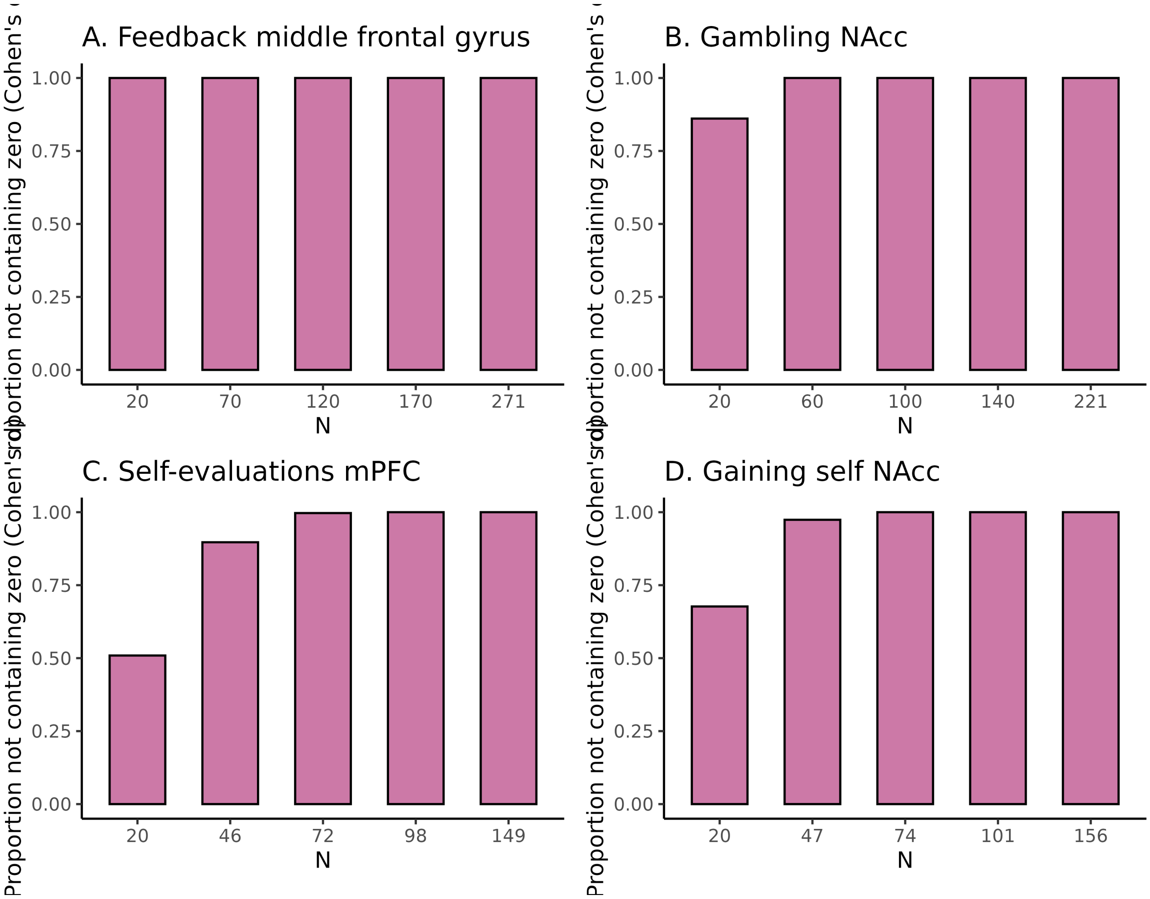

Figure 2: For each task, for five different sample sizes (starting with $N = 20$, then 1/5th parts of the total dataset), the proportion of intervals not containing the value 0 is plotted in reddish purple.

Efron, Bradley, and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. New York: Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593.

Funder, David C., and Daniel J. Ozer. 2019. “Evaluating Effect Size in Psychological Research: Sense and Nonsense.” *Advances in Methods and Practices in Psychological Science* 2 (2): 156–68. https://doi.org/10.1177/2515245919847202.

Gelman, Andrew, and John Carlin. 2014. “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” *Perspectives on Psychological Science* 9 (6): 641–51. https://doi.org/10.1177/1745691614551642.

Gignac, Gilles E., and Eva T. Szodorai. 2016. “Effect Size Guidelines for Individual Differences Researchers.” *Personality and Individual Differences* 102 (November): 74–78. https://doi.org/10.1016/j.paid.2016.06.069.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Klapwijk, Eduard T., Wouter van den Bos, Christian K. Tamnes, Nora M. Raschle, and Kathryn L. Mills. 2021. “Opportunities for Increased Reproducibility and Replicability of Developmental Neuroimaging.” *Developmental Cognitive Neuroscience* 47 (February): 100902. https://doi.org/10.1016/j.dcn.2020.100902.

Klapwijk, Eduard, Herbert Hoijtink, and Joran Jongerling. 2024. *neuroUp: Plan Sample Size for fMRI Regions of Interest Research Using Bayesian Updating*. Zenodo. https://doi.org/10.5281/zenodo.11526169.

Marek, Scott, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, et al. 2022. “Reproducible Brain-Wide Association Studies Require Thousands of Individuals.” *Nature* 603 (7902): 654–60. https://doi.org/10.1038/s41586-022-04492-9.

Maxwell, Scott E. 2004. “The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies.” *Psychological Methods* 9 (2): 147–63. https://doi.org/10.1037/1082-989X.9.2.147.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. “A Manifesto for Reproducible Science.” *Nature Human Behaviour* 1 (1): 0021. https://doi.org/10.1038/s41562-016-0021.

Nord, Camilla L., Vincent Valton, John Wood, and Jonathan P. Roiser. 2017. “Power-up: A Reanalysis of ’Power Failure’ in Neuroscience Using Mixture Modeling.” *Journal of Neuroscience* 37 (34): 8051–61. https://doi.org/10.1523/JNEUROSCI.3592-16.2017.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716.

Peters, Sabine, and Eveline A. Crone. 2017. “Increased Striatal Activity in Adolescence Benefits Learning.” *Nature Communications* 8 (1): 1983. https://doi.org/10.1038/s41467-017-02174-z.

Poldrack, R. A., C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafo, T. E. Nichols, J. B. Poline, E. Vul, and T. 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.

Rouder, Jeffrey N. 2014. "Optional Stopping: No Problem for Bayesians." *Psychonomic Bulletin & Review* 21 (2): 301–8. https://doi.org/10.3758/s13423-014-0595-4.

Satterthwaite, Theodore D., John J. Connolly, Kosha Ruparel, Monica E. Calkins, Chad Jackson, Mark A. Elliott, David R. Roalf, et al. 2016. "The Philadelphia Neurodevelopmental Cohort: A Publicly Available Resource for the Study of Normal and Abnormal Brain Development in Youth." *NeuroImage*, Sharing the wealth: Brain Imaging Repositories in 2015, 124 (January): 1115–19. https://doi.org/10.1016/j.neuroimage.2015.03.056.

Schumann, G., E. Loth, T. Banaschewski, A. Barbot, G. Barker, C. Büchel, P. J. Conrod, et al. 2010. "The IMAGEN Study: Reinforcement-Related Behaviour in Normal Brain Function and Psychopathology." *Molecular Psychiatry* 15 (12): 1128–39. https://doi.org/10.1038/mp.2010.4.

Somerville, Leah H., Susan Y. Bookheimer, Randy L. Buckner, Gregory C. Burgess, Sandra W. Curtiss, Mirella Dapretto, Jennifer Stine Elam, et al. 2018. "The Lifespan Human Connectome Project in Development: A Large-Scale Study of Brain Connectivity Development in 5–21 Year Olds." *NeuroImage* 183 (December): 456–68. https://doi.org/10.1016/j.neuroimage.2018.08.050.

Spaans, Jochem, Sabine Peters, Andrik Becht, Renske van der Cruijsen, Suzanne van de Groep, and Eveline A. Crone. 2023. "Longitudinal Neural and Behavioral Trajectories of Charity Contributions Across Adolescence." *Journal of Research on Adolescence* 33 (2): 480–95. https://doi.org/10.1111/jora.12820.

Szucs, Denes, and John P. A. Ioannidis. 2017. "Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature." *PLOS Biology* 15 (3): e2000797. https://doi.org/10.1371/journal.pbio.2000797.

Turner, Benjamin O., Erick J. Paul, Michael B. Miller, and Aron K. Barbey. 2018. "Small Sample Sizes Reduce the Replicability of Task-Based fMRI Studies." *Communications Biology* 1 (1). https://doi.org/10.1038/s42003-018-0073-z.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." *Frontiers in Psychology* 7. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01832.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Yarkoni, Tal. 2009. "Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul Et Al. (2009)." *Perspectives on Psychological Science* 4 (3): 294–98. https://doi.org/10.1111/j.1745-6924.2009.01127.x.