Reproducible between-person brain-behavior associations do not always require thousands of individuals

Colin G. DeYoung,[1] Tyler A. Sassenberg,[1] Rany Abend,[2] Timothy A. Allen,[3] Roger E. Beaty,[4] Mark A. Bellgrove,[5] Scott D. Blain,[6] Danilo Bzdok,[7,8] Robert S. Chavez,[9] Stephen A. Engel,[1] Ma Feilong,[10] Alexander Fornito,[5] Erhan Genç,[11] Vina Goghari,[12] Rachael G. Grazioplene,[13] Jamie L. Hanson,[14] James V. Haxby,[9] Kirsten Hilger,[15] Philipp Homan,[16] Keanan Joyner,[17] Antonia N. Kaczkurkin,[18] Robert D. Latzman,[19] Elizabeth A. Martin,[20] Luca Passamonti,[21,22] Alan D. Pickering,[23] Adam Safron,[24] Michelle N. Servaas,[25] Luke D. Smillie,[26] R. Nathan Spreng,[27] Jeggan Tiego,[5] Essi Viding,[28] Jan Wacker[29]

[1]Department of Psychology, University of Minnesota, Minneapolis, MN, USA. CGD: cdeyoung@umn.edu; TAS: sasse025@umn.edu; SAE: engel@umn.edu

[2]National Institute of Mental Health, Bethesda, MD, USA. rany.abend@nih.gov

[3]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. timothy.a.allen@gmail.com

[4]Department of Psychology, Pennsylvania State University, University Park, PA, USA. rebeaty@psu.edu

[5]Turner Institute for Brain and Mental Health, Monash Biomedical Imaging, and School of Psychological Sciences, Monash University, Clayton, Australia. MAB: mark.bellgrove@monash.edu; AF: alex.fornito@monash.edu; JT: jeggan.tiego@monash.edu

[6]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA. scottdougblain@gmail.com

[7]Department of Biomedical Engineering, McConnell Brain Imaging Centre, Montreal. Neurological Institute (MNI), Faculty of Medicine, School of Computer Science, McGill University, Montreal, Canada. danilobzdok@gmail.com

[8]Mila - Quebec Artificial Intelligence Institute, Montreal, Canada.

[9]Department of Psychology, University of Oregon, Eugene, OR, USA. rchavez@uoregon.edu

[10]Department of Psychology, Dartmouth College, Hanover, NH, USA. MF: feilong.ma@dartmouth.edu; JVH: james.v.haxby@dartmouth.edu

[11]Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany. genc@ifado.de

[12]Department of Psychiatry, University of Toronto, Toronto, Canada. vina.goghari@utoronto.ca

[13]Department of Psychiatry, Yale University, New Haven, CT, USA. rachael.grazioplene@gmail.com

[14]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA. jamielh@pitt.edu

[15]Department of Psychology, Universität Würzburg, Würzburg, Germany. kirsten.hilger@uni-wuerzburg.de

[16]Department of Psychiatry, Psychotherapy and Psychosomatics, University of Zürich, Zürich, Switzerland. philipp.homan@bli.uzh.ch

[17]Department of Psychology, University of California, Berkeley, CA, USA. kjoyner@berkeley.edu

[18]Department of Psychology, Vanderbilt University, Nashville, TN, USA. antonia.kaczkurkin@Vanderbilt.Edu

[19]Takeda, Cambridge, USA. rlatzman@gmail.com

[20]Department of Psychology, University of California, Irvine, CA, USA. emartin8@uci.edu

[21]Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. passamontiluca@gmail.com

[22]Institute of Bioimaging and Molecular Physiology, Consiglio Nazionale delle Ricerche, Rome, Italy.

[23]Department of Psychology, Goldsmiths, University of London, London, UK. A.Pickering@gold.ac.uk

[24]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. asafron@gmail.com

[25]Faculty of Medical Sciences, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. m.n.servaas@umcg.nl

[26]Department of Psychology, University of Melbourne, Melbourne, Australia. lsmillie@unimelb.edu.au

[27]Department of Neurology and Neurosurgery, McGill University, Montreal, Canada. nathan.spreng@gmail.com

[28]Clinical, Educational, and Health Psychology Research Department, University College London, London, UK. essividing@gmail.com

[29]Institute of Psychology, Universität Hamburg, Hamburg, Germany. jan.wacker@uni-hamburg.de

Arising from: Marek et al., Reproducible brain-wide association studies require thousands of individuals. https://doi.org/10.1038/s41586-022-04492-9 (2022)

Marek et al. analyzed three very large magnetic resonance imaging (MRI) datasets and concluded that thousands of participants are necessary to ensure replicable results in "brain-wide associations studies," which they defined as "studies of the associations between common inter-individual variability in human brain structure/function and cognition or psychiatric symptomatology."[1] This conclusion overgeneralizes the implications of their findings and is likely to have an unwarranted chilling effect on neuroimaging research focused on individual differences, preventing good research with samples in the hundreds from being funded and conducted. To fend off these negative consequences, we explain why their conclusion is not fully justified, discuss methods that can yield larger effects, and suggest practical guidelines for sample size, recognizing the potential utility of samples in the hundreds.

How large samples need to be depends primarily on effect size. The central conclusion of Marek et al. boils down to a claim about the expectable range of effect sizes in cross-sectional studies of associations between behavioral measures and variables derived from MRI data. If all expectable effect sizes are as low as they claim (they found the largest 1% of replicable univariate effects to be between $|r| = .06$ and .16), then samples in the thousands could be necessary for adequate statistical power, especially when conducting multiple statistical tests.

The median reported effect size in meta-analyses of research on behavioral individual differences is about $|r| = .20$.[2,3] Below, we review research demonstrating that effect sizes for between-person brain-behavior associations, when using appropriate methods, can be this large or larger. To detect a Pearson correlation of .20 at $p < .05$ with 80% power requires a sample of just under 200 participants. Thus, we suggest 200 as a more reasonable minimum sample size for correlational neuroimaging research, provided there is reason to expect an effect size of at least .20 and the focal analysis involves only a single statistical test. Larger samples are often

desirable because multiple tests may need to be conducted and researchers may wish to detect smaller effects. However, researchers should not be afraid to use samples in the hundreds rather than the thousands, if they have evidence that allows them to anticipate sufficient effect sizes.

Expected effect sizes cannot be generalized from one set of methods to all others. In both their univariate and multivariate analyses, Marek et al. drew conclusions based on a narrow set of methods, relative to the wide array of possible methods. We identify three ways in which different methods can produce larger effect sizes for brain-behavior associations, thereby allowing detection of replicable findings in samples smaller than 1000.

(1) *Theoretical matching between behavioral and neural variables.* Neither resting-state functional MRI (fMRI) nor structural MRI directly assess an aspect of brain function that is transparently relevant to any specific behavioral characteristic. Task-based fMRI, however, can induce brain states that correspond more clearly to the behavioral measure in question. Task-based brain-behavior associations are consistently stronger than resting-state based associations in cross-validation studies of very large samples, often with effects greater than $r = .20$.[4-7] Marek et al. acknowledged that fMRI task data may yield larger effect sizes, but they also improperly discounted the one example of such an effect size that they reported from their analyses. In Extended Data Figure 3, they reported that the correlation between cognitive ability and activation of the dorsal attention network during a working memory task was .34. However, they dismissed this finding by mischaracterizing working memory performance as a "confound" that needs to be controlled (yielding a much smaller correlation of .14). The plausible causal arrangement of the three variables—working-memory-related neural activity, working memory performance, and general cognitive ability—is not one of confounding. Rather, working memory performance (which is a persistent trait strongly correlated with general cognitive ability[8]) should

act as a mediator between neural activity and general cognitive ability. Thus, the control was inappropriate and the correct and theoretically meaningful correlation was .34. Another study of the same sample, using multivariate instead of univariate methods, also found neural activation during various tasks to predict cognitive ability with correlations around .30.[7] Similarly high multivariate correlations with cognitive ability were found for task-related functional connectivity in 1858 participants from a different sample.[4]

(2) *Individualized localization.* Although the methods used by Marek et al. are common in the field, their effect sizes are unlikely to be fully generalizable because they do not reflect best practices for assessing neural variables across participants. They analyzed individual vertices or brain regions of interest from standard atlases without any adjustments for the fact that functional organization of the brain, especially in cortex, differs from person to person in relation to neuroanatomical landmarks. Simply aligning neuroanatomy and then comparing individuals based on anatomical locations ensures that the properties of the vertices or regions studied will not be fully comparable across participants, thus increasing noise and reducing effect sizes.

Fortunately, a number of techniques exist for adjusting neural measurements so as to compare functionally equivalent vertices, voxels, or regions across participants.[5,9-11] One strategy involves iterative Bayesian realignment of standard atlases, as employed in *group prior individualized parcellation*[9] and *multi-session hierarchical Bayesian modeling* (MS-HBM).[11] In 1094 participants, Kong et al. showed that multivariate patterns of resting functional connectivity predicted 58 behavioral variables more strongly after applying MS-HBM to a standard parcellation atlas than before doing so.[11] Further, eight of the resulting correlations were larger than .20.

Another promising technique is *hyperalignment*, which addresses the fact that, even within a functional brain region, the same information is processed in different voxels in different brains. Hyperalignment identifies sets of voxels with similar patterns of neural activity for each participant and treats them as the relevant neural unit of analysis. In the same sample in which Marek et al. reported correlations with working-memory-related activation, Feilong et al. were able to predict cognitive ability using estimates of functional connectivity based on hyperalignment, with average multivariate effect sizes of $r = .53$ for task data and $r = .44$ for resting-state, across all brain regions.[5]

(3) *Improved measurement.* In both the neural and behavioral domains, observed effect sizes are limited by the reliability and validity of the variables under investigation. Although Marek et al. claim adequate reliability for their measures, analysis of the same data using item response theory has shown inadequate reliability for their measure of psychopathology.[13] In both neural and behavioral data, reliability and validity may be improved by (among other strategies) structural equation modeling with latent variables, which remove measurement error from the constructs of interest.

In conclusion, the relatively large effect sizes that we cited here provide proof of principle for the existence of effects larger than those highlighted by Marek et al. With larger effect sizes, smaller samples are viable. Even if multivariate effects often become smaller when replicated in other samples, some of them are stable and larger than $|r| = .20$ when validated out-of-sample.[13] Multivariate methods are powerful tools for identifying effects in samples less than 1000.[14,15] Blanket dismissal of neuroimaging research on individual differences in samples ranging from 200–1000 participants would be a grave mistake, squandering valuable research opportunities and preventing many labs from conducting important neuroimaging research, thus

rendering the field more exclusionary and greatly reducing avenues for creativity, diversity, and discovery.

References

1. Marek, S., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654-660.
2. Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74–78.
3. Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78–79.
4. Chen, J., et al. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature Communications*, *13*(1), 1-17.
5. Feilong, M., Guntupalli, J. S., & Haxby, J. V. (2021). The neural basis of intelligence in fine-grained cortical topographies. *eLife*, *10*, e64058.
6. Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature communications*, *9*(1), 2807.
7. Sripada, C., Angstadt, M., Rutherford, S., Taxali, A., & Shedden, K. (2020). Toward a "treadmill test" for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping*, *41*(12), 3186-3197.
8. Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151-177.

9. Chong, M., et al. (2017). Individual parcellation of resting fMRI with a group functional connectivity prior. *NeuroImage*, *156*, 87-100.
10. Gordon, E. M., Laumann, T. O., Adeyemo, B., & Petersen, S. E. (2017). Individual variability of the system-level organization of the human brain. *Cerebral Cortex*, *27*(1), 386–399.
11. Kong, R., et al. (2021). Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex,* 31, 4477–4500.
12. Tiego, J., & Fornito, A. (2022). Putting behaviour back into brain-behaviour correlation analyses - Commentary on Marek et al. (2022). Preprint at http://osf.io/srcbm
13. Chavez, R. S., Tovar, D. T., Stendel, M. S., & Guthrie, T. D. (2022). Generalizing effects of frontostriatal structural connectivity on self-esteem using predictive modeling. *Cortex*, *146*, 66-73.
14. Bzdok, D., Engemann, D., & Thirion, B. (2020). Inference and prediction diverge in biomedicine. *Patterns*, *1*(8), 100119.
15. Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M., & Kennedy, K. M. (2021). Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Human Brain Mapping*, *42*(1), 204-219.

Data Availability Statement: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Author Contributions: Conception: CGD, JW, RNS, and SAE conceived of the commentary. CGD wrote the initial draft. All authors were involved in editing and revising.

Competing Interests Statement: The authors declare no competing interests.