How sample size influences the reproducibility of task-based fMRI

Erick. J. Paul[1,8], Benjamin O. Turner[2,8], Michael B. Miller[2], Aron K. Barbey[1,3,4,5,6,7]
[1]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign
[2]Department of Psychological & Brain Sciences, University of California, Santa Barbara
[3]Department of Psychology, University of Illinois Urbana-Champaign
[4]Neuroscience Program, University of Illinois Urbana-Champaign
[5]Department of Bioengineering, University of Illinois Urbana-Champaign
[6]Department of Internal Medicine, University of Illinois Urbana-Champaign
[7]Carle R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign
[8]These authors contributed equally.

Despite a growing body of research suggesting that task-based functional magnetic resonance imaging (fMRI) studies often suffer from a lack of statistical power due to too-small samples, the proliferation of such underpowered studies continues unabated. Using large independent samples across four distinct tasks, we demonstrate the impact of sample size on reproducibility, assessed at different levels of analysis relevant to fMRI researchers. We find that typical sample sizes produce results that have a low degree of reproducibility, and even samples much larger than typical (e.g., $N = 100$) produce results that are far from perfectly reproducible. Thus, our results join the existing line of work advocating for larger sample sizes. Moreover, because we test sample sizes over a fairly large range and use intuitive metrics of reproducibility, our hope is that our results help catalyze a major shift in how task-based fMRI research is carried out across the entire field.

Keywords: reliability, statistical parametric mapping, replication

Introduction

Recent years have seen the arrival of a "reproducibility crisis" in science, both in science at large (Baker et al., 2016; Munafò et al., 2017; "Replication studies offer much more than technical details", 2017), and perhaps even more acutely in the psychological sciences (Open Science Collaboration, 2015). Some of the reasons behind this crisis—including flawed statistical procedures, career incentive structures that emphasize rapid production of "splashy" (i.e., unlikely) results while punishing "failed" studies, and biases inherent in the publication system—have been articulated carefully in previous work, again both generally (Szucs, 2016; Barnes, Tobin, Johnston, MacKenzie, & Taglang, 2016; Wicherts et al., 2016), and for fMRI in particular (Carp, 2012; Button et al., 2013; Poldrack et al., 2017; Szucs & Ioannidis, 2017). Among these problems, the most frequently identified, and possibly the most easily remedied, is lack of statistical power due to too-small samples. Indeed, the field of fMRI has seen recommendations against large samples (e.g., Friston, 2012; cf. Ingre, 2013), and even when larger sample sizes are acknowledged as desirable, what constitutes "large enough" has often been an ad-hoc process of developing unempirical rules of thumb.

Of course, this lack of power is driven in large part by the great expense associated with collecting fMRI data (Mumford & Nichols, 2008). Even relatively small studies can cost several tens of thousands of dollars, and the funding system throughout much of the world is not generally set up to enable the routine collection of large (e.g., $N > 100$) samples. However, aside from these financial considerations, there are two other reasons researchers may persist in collecting small samples. The first is that while tools exist that allow researchers to do prospective power analyses for fMRI studies (Mumford & Nichols, 2008; Durnez et al., 2016), researchers may struggle to understand these tools, because defining power in an fMRI context involving hundreds of thousands of statistical tests is conceptually distant from defining power in a typical behavioral context, where there might be on the order of ten such tests. Relatedly, meaningfully defining "effect size" is conceptually straightforward in a behavioral context, but much less so in an fMRI context.

The second possible non-financial reason that researchers continue using small samples is because a number of studies have shown that fMRI has reasonably good test-retest reliability (Bennett & Miller, 2010; Gonzalez-Castillo & Talavage, 2011; Plichta et al., 2012; Bennett & Miller, 2013). It is possible that researchers take this to mean that large samples are not necessary, particularly if the researcher uses design optimization approaches to increase power at the in-

dividual level (Liu, Frank, Wong, & Buxton, 2001; Wager & Nichols, 2003; Liu & Frank, 2004). However, test-retest reliability is not only not synonymous with replicability, but it is in some ways antithetical. This is because typical measures of test-retest reliability, e.g. the intra-class correlation (ICC), rely on variability across individuals. However, replicability is reduced by individual variability, particularly with small samples. While it is true that a measure with low test-retest reliability will have low replicability (in the limit, all individual maps are pure noise, and if there are suprathreshold voxels in the group average map, they likewise represent non-reproducible noise), it does not follow that high test-retest reliability guarantees replicability at the level of group-average maps. Nor is it the case that variability between individuals in terms of brain activity is so minor that we can disregard it when considering the relationship between test-retest reliability and replicability; on the contrary, research has demonstrated that variability between individuals can swamp group-average task-related signal (Miller et al., 2002, 2009; Miller, Donovan, Bennett, Aminoff, & Mayer, 2012; Gabrieli, Ghosh, & Whitfield-Gabrieli, 2015).

Our goal in the present study is to provide empirical estimates of fMRI's replicability in terms of the levels of results that are useful in the field (i.e., multi-voxel patterns or cluster-based results, rather than, e.g., peak t-statistic values). Our specific focus is on the role of sample size (i.e., number of participants) on replicability, although we do examine the influence of other factors that should affect replicability, including design power (Turner & Miller, 2013). We also emphasize that our results, far from being relevant only to researchers whose specific interest is in studying reproducibility (e.g., Evans, 2017), are applicable to all researchers who are interested in using fMRI to produce valid and meaningful neuroscientific discoveries. In fact, we use $N \simeq 30$ as our standard for a "typical" fMRI sample size, which is in line with empirical estimates by Szucs and Ioannidis (2017) (median sample size of fMRI studies in 2015 = 28.5) and Poldrack et al. (2017) (75th percentile of sample size in cognitive neuroscience journals published between 2011–2014 = 28). To preview our results, we provide an easily-interpretable demonstration of the facts laid out by Button et al. (2013) and Szucs and Ioannidis (2017): An examination of multiple measures of reproducibility computed across multiple levels of analysis demonstrates that replicability of fMRI studies with a sample size anywhere near 30 is strikingly low.

Method

We carried out a series of analyses across four distinct tasks. Because these analyses had the same form across all four tasks, we describe here the details of those analyses, and leave the description of the details specific to each task to the Supplemental Materials. We refer to the four tasks throughout this report as A, B, C, and D, because our interest is not in the identity of these tasks *per se*.

Participants

Participants were recruited from the Urbana-Champaign community as part of two separate intervention studies, each of which included a pre-intervention MRI session with two different fMRI tasks (for a total of four fMRI tasks). Both studies were approved by the University of Illinois Urbana-Champaign Institutional Review Board; all participants in both intervention experiments provided informed consent. All participants were right-handed, had normal or corrected-to-normal vision without color blindness, reported no previous neurological disorders, injuries, or surgeries, reported no medications affecting central nervous system function, were not pregnant, had no head injuries or loss of consciousness in the past two years, and were proficient in English. All participants received monetary compensation for participation. Only data provided at the pre-intervention time point (i.e., prior to the start of any intervention or experimental conditions) are included in the present analyses.

A total of 301 participants were recruited for and provided data in the first intervention study (Study 1). For the two fMRI tasks A and B, an identical set of 279 participants had complete data in both and are included in all analyses.

A total of 227 participants were recruited for and provided data in the second intervention study (Study 2). Task C includes a sample of 214 participants with complete data, and Task D includes 200 participants (of the 214 included in Task C) with complete data.

Scanning procedure

All participants in both Studies 1 and 2 were scanned on the same Siemens 3T Magnetom Trio. Study 1 participants were scanned with a 32-channel head coil; Study 2 participants were scanned with a 12-channel head coil. High resolution anatomical data were obtained using a high resolution 3D structural MPRAGE scan: 0.9 mm isotropic, TR = 1900 ms, TI = 900 ms, TE = 2.32 ms, with a GRAPPA acceleration factor of 2. Functional MRI BOLD data were collected using the Siemens echo-planar imaging sequence. Tasks A, B, and D used the following parameters: TR = 2000 ms, TE = 25 ms, flip angle = 90°, 92×92 matrix with 2.5 mm in-plane resolution, 38 slices parallel to AC-PC with a 3.0 mm slice thickness and 10% slice gap. Task C used the same parameters, with the exception of the following: TR = 2360 ms, 45 slices with a 2.5 mm slice thickness. The number of repetitions varied for each task depending on the task duration (see Supplemental Materials for details). Finally, a gradient field map was collected for use in B0 unwarping matching the EPI parameters.

Preprocessing

Every run from each task was preprocessed identically using FSL's (http://www.fmrib.ox.ac.uk/fsl) FEAT (FMRI Expert Analysis Tool, version 6.00) software package. Preprocessing included motion correction using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), BET brain extraction (Smith, 2002), spatial smoothing with a 6 mm FWHM kernel, grand-mean intensity normalization, prewhitening with the FILM tool (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) and a high pass filter with a cutoff of (1/90) Hz. EPI images were additionally unwarped using the gradient field maps collected with the functional runs. The high-resolution structural scan was registered to the MNI152-T1-2mm standard brain via FLIRT (Jenkinson & Smith, 2001; Jenkinson et al., 2002) and further refined using the non-linear FNIRT tool (8mm warp resolution; Andersson, Jenkinson, Smith, et al., 2007). Transformation of each functional scan to the MNI standard brain was accomplished using a two-step process to improve alignment first by registering each EPI to the high-resolution structural scan with the FSL BBR tool (Greve & Fischl, 2009), and then applying the non-linear warp generated from the high-resolution scan to the functional scan.

GLM analysis

For a complete description of each task, task events, and contrasts, see Supplemental Materials. Briefly, Task A included 7 experimental events; Task B included 10 events; Task C included 7 events; Task D included 4 events. Predicted BOLD signals were generated for each event via convolution with a double gamma HRF (phase $= 0$). Six regressors derived from the motion parameters were included as regressors of no interest in each low-level model to mitigate the effects of motion in the data. The temporal derivative of each event was also included and the same temporal filtering that was applied to the preprocessed data was also applied to the model. A primary contrast of interest was identified for each task, defined by the cognitive effect that the task was designed to capture (i.e., the contrast an experimenter running any of these particular tasks would be primarily interested in). The contrast of interest was estimated in each subject in a mid-level analysis by combining all runs in a fixed-effects model. Following that, group-level statistical results for each task/contrast were generated using a mixed-effects model via FSL's FLAME1 tool (Woolrich, 2008).

Pseudo-replicate analysis

To estimate the reproducibility of group-level results, we took the following approach. First, we split our full sample of N participants into two randomized, non-overlapping sets ("P" and "Q") of length $N/2$. Next, we chose a sample size $k \in \{16, 25, 36, 49, 64, 81, 100(, 121)\}$ for which we sought to estimate the reproducibility, and used FSL's FLAME1 tool to generate group-level statistical maps using the first k participants in both groups P and Q. Then, for each of a number of similarity measures, we computed the similarity between the P and Q group-level maps. Finally, we repeated the preceding steps across all in-range values of k, and for 500 random sorts in groups P and Q.

This same process was carried out for every task; for Tasks C and D, all sorts were done independently, while for Tasks A and B (which comprised an identical set of participants), the same 500 sorts were applied to both tasks. For the purposes of presentation, we show the average reproducibility estimate across all four tasks for each sample size, along with the average within-task (and within-sample size) standard deviation, though we also include the curves for each task. Although we present error bars for all of our analyses, note that, as with all resampling-based analysis methods, our results suffer from complex interdependence that makes it difficult to draw strong inferences about differences between tasks. That is, the variance among the 500 simulated replications of a given task in our approach may underestimate the variance that would be observed given 500 true, completely independent replications of the task. Moreover, there is no analytic solution that would let us correct for this underestimation, if in fact it exists. Therefore, all error bars should be interpreted as being qualitative or illustrative, rather than as guides for whether differences are significant. To that end, we use standard deviations rather than standard errors or confidence intervals in our presentation of the results.

Similarity statistics

The similarity statistics that we used to operationalize reproducibility were chosen to reflect different levels of focus. Broadly, there were three levels, which from most to least granular were voxel, cluster, and peak. We describe the measure(s) associated with each level in turn below. Throughout our analyses, we present results in an "exact replication" frame—that is, our results provide an empirical demonstration of what a researcher could expect if she were to re-run a study exactly, down to the sample size of the original study. Our gold standard would be to present results that reflect how well a study's results capture "ground truth" as a function of sample size. Unfortunately, as is generally the case, the ground truth for the experimental contrasts we have included here is unknown.

Previous investigations in a similar vein have used either a meta-analytic approach or results from "large-enough" samples to approximate ground truth. However, meta-analyses suffer from well-established biases against small (but putatively nonetheless significant) results, and are moreover ill-suited to address some of the levels we focus on here. Likewise, to preview our results, although we have access to large samples by the standards of many neuroimaging studies, they

are not large enough to establish a reliable ground truth. More to the point, because of differences between tasks in terms of power and maximum available sample sizes, these ground truth maps would reflect different levels of "truthiness" across tasks, which would further confuse interpretation of these results. However, we do use results from the full sample in our "voxel-level (thresholded)" analyses, as described in more detail below.

We also describe the method associated with each measure for determining the measure's null distribution. To aid readers in placing our results in the appropriate context, we generated results expected under two distinct null hypotheses for each of our measures. In every case, one of these null hypotheses was simple in both concept and implementation, while the other was more complex and meant to represent a more realistic null hypothesis. We refer to these as the "simple" and "realistic" null hypotheses, but the terms "weak" and "strong" or "liberal" and "conservative" could also be applied. While we based our realistic null approaches on reasonable assumptions, the simple null results serve as an absolute lower bound for those readers who feel that our realistic null approaches represent too high a bar. In the main text, we compromise and present results averaged across the two null approaches (which we refer to as the "hybrid" null); individual null results are available in the Supplemental Materials.

Voxel-level reproducibility (intensity). Arguably, the goal in fMRI is to accurately capture the activity in every single voxel. Indeed, many analysis methods are predicated on just such an assumption (e.g., MVPA, RSA, or encoding models), and many techniques for improving data acquisition or preprocessing are aimed at getting ever-finer spatial resolution (which we presume would be wasted effort if researchers' goal was merely to approximate the spatial location of activity, or equivalently, the activity associated with a given location). Therefore, the first level of reproducibility on which we focused was the reproducibility of voxel-wise intensities.

To quantify similarity, we used the Pearson correlation, which ranges from -1 (inverse SPMs, invariant to scale) to 1 (identical SPMs, invariant to scale). Although the Pearson correlation is not a measure of reliability, it does give us an indication of how similar the between-voxel patterns of activity are across SPMs. (Our results are qualitatively unchanged if a measure not based on covariance, e.g., eta^2, is used instead of the Pearson correlation.) To generate this measure, we computed the similarity between the vectorized unthresholded group-level SPMs, after applying a common mask to remove voxels that were zero (i.e., outside of the group FOV) in either SPM.

The simple null distribution for both metrics were constructed by generating SPMs of white noise spatially smoothed to match the observed smoothness in our real SPMs, and rescaled to equate the robust min and max (i.e., 2$^{\text{nd}}$ and 98$^{\text{th}}$ percentile, respectively). For each task and sample size, we generated the observed histogram of estimated FWHMs (using FSL's `smoothest` command) as well as observed histograms of robust mins and maxes. We then parameterized these histograms and drew 1000 samples from the resulting parametric normal distributions. Finally, we generated 1000 maps of pure $\mathcal{N}(0,1)$ noise, smoothed each map with the corresponding sampled FWHM (using FSL's `fslmaths` utilities) and rescaled to match the sampled robust min and max.

The realistic null distribution for both of these was constructed to reflect the strict interpretation of "reproducible" we stated for this level. In particular, we generated a set of R null maps separately for every task, sample size, and bootstrap iteration (with the caveat that Tasks A and B relied on the same set of null maps). The null maps were generated using a novel "voxel drift" resampling algorithm: for a given SPM, every voxel's position was resampled probabilistically according to that SPM's empirical FWHM (with the caveat that every voxel had to shift from its original position). Then, given this set of R null maps and the counterpart true map (i.e., if Q had been chosen for this voxel drift resampling procedure, the P map would be left unchanged), each statistic was computed for the comparison of the true map with all R null maps. Finally, to be able to give a "chance" performance curve, we recorded the 95$^{\text{th}}$ percentile of each statistic across these R calculations. This null hypothesis can be stated as, "The location of a particular voxel is random to within roughly FWHM mm."

Voxel-level reproducibility (thresholded). Without abandoning the notion of describing reproducibility at the voxel level, it is nonetheless possible to relax the definition of what is being reproduced somewhat—i.e., from raw intensity value to a binary "active"/"inactive" classification. To this end, we carried out a second set of analyses at the voxel level, using thresholded, binarized maps. As alluded to earlier, we used full-sample results in these analyses. Specifically, we thresholded each of the full-sample SPMs at liberal and conservative thresholds using FSL's cluster-based thresholding (see next section for additional details on cluster-based thresholding), and used these thresholded maps in order to estimate the "true" proportion of voxels that should be suprathreshold for each task.

With these per-task proportions suprathreshold, we simply applied proportion-based thresholding of the group-level SPMs (two-tailed) in order to match the full sample proportion suprathreshold. Conceptually, this is distinct from the cluster-based thresholding used in the subsequent sections, in that the voxels which end up suprathreshold are not guaranteed to meet any particular cutoff for significance, either at the voxel level or familywise. Thus, the quantity that is held constant across group-level maps P and Q is not the theoret-

ical Type I or II error rates of each map, but simply the number of suprathreshold voxels. Our metric of reproducibility for these thresholded maps was the Jaccard statistic, which is simply the ratio of the intersection of a pair of thresholded maps divided by their union. This statistic ranges from 0 (no overlap) to 1 (perfect overlap).

The simple and realistic null results were generated using the same approaches outlined for the intensity-based voxel-level analyses, with the added steps of thresholding (two-tailed) the null maps at the same target proportion and computing the Jaccard overlap between the pair of one null and one true map. As above, this procedure was repeated 1000 times and the 95th percentile was taken.

Cluster-level reproducibility. While the ultimate or idealized goal of fMRI would seem to be voxel-level reproducibility, the common currency of today's analytic landscape is generally the cluster (or as a special case, the peak; see next section). Therefore, the second level of reproducibility on which we focused was at the cluster level. Here, we chose to focus simply on the binary distinction between sub- and supra-threshold that forms the basis of cluster-based approaches (along with others). Although cluster-based approaches are widely used, it is less clear exactly what it means to reproduce a cluster. Existing methods for conducting inferential statistics on clusters (e.g., Gaussian random field theory, Worsley, Taylor, Tomaiuolo, & Lerch, 2004; or permutation, Nichols & Holmes, 2002) refer to the null probability of observing a cluster of a given size (or possibly mass; Zhang, Nichols, & Johnson, 2009) conditioned on an initial threshold level, but do not address the question of exactly where this cluster appears.

Certainly, the spatial resolution at the cluster-level is coarser than at the voxel-level—researchers generally do not expect that every single supra-threshold voxel in a given cluster would be supra-threshold under replication, and likewise with sub-threshold voxels. Durnez, Moerkerke, and Nichols (2014), from which we take inspiration for our peak-based approach, employed a liberal definition in their cluster-based methods: a cluster is "replicated" if a single voxel from a given cluster is suprathreshold in replication. For our application, such a definition is far too generous, so we once again used Jaccard overlap. To generate clusters, we used FSL's cluster-based thresholding on every group-level SPM, once at a liberal threshold and once at a more conservative threshold.

We note as well some researchers might view cluster replication as a question of proximity; although Jaccard overlap is not a measure of proximity, it will generally track with proximity (i.e., as two clusters get closer together, their Jaccard overlap will increase). The exception to this is in the case of clusters which have zero intersection; a proximity-based measure would distinguish between a proximal pair of (non-intersecting) clusters and a distal pair, while both would have a Jaccard overlap of 0. In the interest of simplicity, as well as conceptual rigor when it comes to defining replication, we eschew such proximity-based measures.

Given the "fuzzier" localization inherent at the cluster level, it makes little sense to use the voxel drift resampling procedure we outlined above in this context. On the other hand, completely unconstrained drift (i.e., simply matching the number of supra-threshold voxels) is certainly too lax, and also fails to respect the spatial correlation inherent in cluster-thresholded maps (which will change the shape of the null distribution). Therefore, we used a simple procedure to generate null clusters: for a given task, we constructed the corpus of all clusters from all analyses across the other three tasks. To match a given true image with a corresponding null image, we drew random clusters from this corpus, while imposing the constraint that the proportion of suprathreshold voxels in the null map be within ± 0.01 of the proportion in the corresponding true image. (Note that these clusters are independent, though they may be spatially similar to the extent that these contrasts rely on common processes. Given the ubiquity of subtraction logic in neuroimaging, we view the possibility that these clusters will be similarly located as a feature of this approach, rather than a bug.) The realistic null was computed as the Jaccard overlap between the generated map and the true map. The simple null was computed almost identically to that described in the previous section: each null smoothed map was thresholded (one tailed) to match the proportion of suprathreshold voxels from the corresponding true image, and the Jaccard overlap between the two was computed.

Peak-level reproducibility. The last analysis we report focuses on the level of peaks. Although clusters form the foundation of the majority of thresholding-based analyses used today, these clusters are typically reported simply in terms of the location and intensity of their peaks. In fact, some recent work has developed the statistical framework for understanding the behaviors of peaks, and how this can be used in, e.g., power analyses (Mumford & Nichols, 2008; Durnez et al., 2014). For the present purposes, we do not need to know the distributional characteristics of peaks, nor do we need to use the sophisticated estimation procedures described by Durnez et al. (2014). Therefore, we use the same cluster-extent (Gaussian random field theory) thresholding approach as for the cluster-level analyses. That is, whereas Durnez et al. (2014) use a peak-based secondary threshold when considering peaks as topological features, we use an extent-based secondary threshold. However, as in Durnez et al. (2014), this results in a set of surviving peaks and a set of non-surviving peaks.

In contrast to the previous levels, we consider two types of replication at the peak level. The first is the same as in Durnez et al. (2014): a peak is considered replicated if it is suprathreshold under replication (i.e., part of any surviving

cluster). This is a fairly generous definition of replication, but much less so than their cluster-level approach (i.e., non-zero overlap between clusters). Although we cannot interpret results in terms of false positives or false negatives, we can nonetheless examine the replication success of surviving and non-surviving peaks separately. (We additionally provide an analysis of the relationship between peak z values and reproducibility for surviving peaks in the Supplemental Materials.) Therefore, for each of these two sets, we compute the proportion of peaks in one map that are suprathreshold in the complementary map. And unlike all previous measures, this measure is asymmetric—the proportion of P peaks that are suprathreshold in Q need not be equal to the proportion of Q peaks suprathreshold in P—so we calculated it in both directions and then averaged the results to arrive at the final value.

We used the same two approaches described in the preceding section to generate realistic and simple null distributions. That is, for the realistic null, we used the same maps of independently-generated clusters to determine whether each peak was supra- or sub-threshold. For the simple null, we used the smoothed null maps, thresholded to match proportion, to classify peaks.

Measurables

Our expectation was that sample size would be the largest driver of reproducibility, irrespective of how it was measured. However, we also expected variability between our tasks (which would be unexplainable by k), as well as variability within a task for a given k (which would be unexplainable both by k and by task-level variables). Therefore, we carried out an analysis in an attempt to find other easily-measured variables that might explain these two types of variability. Although our primary goal is descriptive—that is, to identify the relationships present in our data—we used a modeling approach that in principle should allow generalization.

Before we describe this approach in detail, we note that we cannot use standard regression techniques to derive inferential statistics for our regressors, because our observations are non-independent (i.e., the correlation between any P and Q group-level maps reflect contributions from specific participants, all of whom will almost surely be members of other P or Q groups). Moreover, the influence of this non-independence varies across sample sizes, because the average number of participants in common between any two groups across iterations at a sample size of, say, 16, will be much lower than the average number of participants in common between groups at a sample size of 100.

The modeling approach we used included several steps. First, because of the aforementioned fact that non-independence exerts the greatest influence for the largest sample size, we removed the 500 observations from each task corresponding to the largest sample size (either 100 or 121). Next, we removed the task effects from the outcome variable as well as all predictor variables (this was done to simplify interpretability of R^2 values; the Frisch-Waugh-Lovell theorem ensures that including a categorical regressor for task would yield identical beta estimates), z-scored all variables, and fit a full model. We used this fit to identify how many variables to retain (a value we refer to as m below) in subsequent steps by examining the plot of beta estimates for an obvious elbow, and also to establish the common within-task effect of each regressor.

We then used a double-cross-validation approach to verify that the variables we selected generalized across tasks. At the first cross-validation loop, we held out a single task as a test set; at the second cross-validation step, we held out each of the remaining tasks in turn, fit a model on the remaining two tasks with all regressors (demeaned by task), and used the regression estimates to select the m best variables. Next, we combined the three tasks that had not been held out (now without task effects removed), selected the m variables that had occurred most frequently across the second cross-validation loop (using absolute normalized beta value to break ties when necessary; this second CV was required to avoid circularly using the test data for model selection and validation), fit the model including these m variables for the three tasks, and used the estimated betas to predict the (un-task-demeaned) outcome variable using the (un-task-demeaned) m variables from the hold-out task. We compare root mean squared error (RMSE) against the (biased) standard deviation of the outcome variable and the RMSE from a null model including only an intercept term to assess performance. We also examine regressor prevalence and beta estimate stability.

We chose correlation as the outcome measure of interest for our analysis of the influence of various data properties because it is generally smoother (less stochastic) than the other measures. However, we demonstrate in the Supplemental Materials that the models fit to correlation generalize to the other measures, so the qualitative pattern of our results does not depend on the choice of outcome measure.

Results

We present the result from each of the three levels of analysis described in the Methods—voxel, cluster, and peak—in separate sections below. Each section includes the true observed results for the measure used at that level in terms of the impact of sample size and task on that measure, as well as null results from the hybrid null (i.e., the average of the simple and realistic null methods). In a separate section, we explore the relationship between various measurable properties of the data and the voxel-level reproducibility results.

For all Figures throughout the first three sections below, note that we plot results as lines for clarity, but computed our

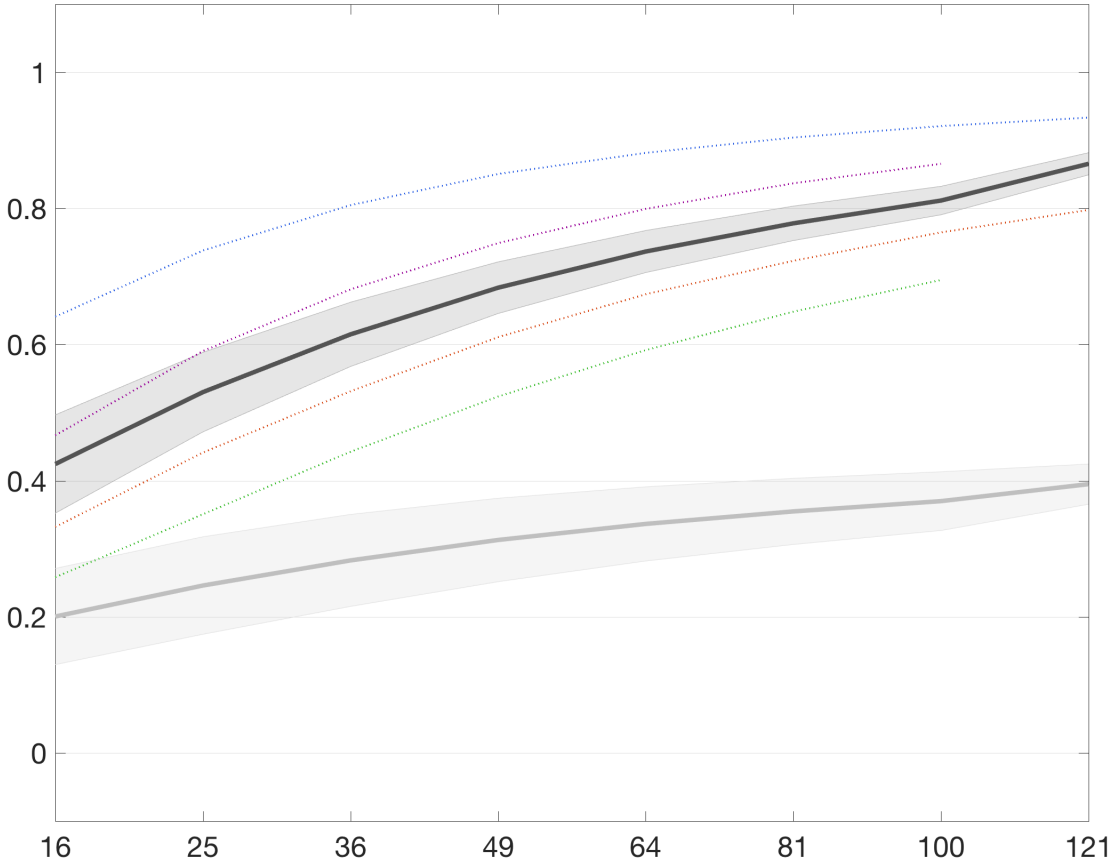


Figure 1. Replicability results for voxel-level (unthresholded) analyses. Average observed (± 1 standard deviation) shown in black (dark gray); average hybrid null (± 1 standard deviation) shown in medium gray (light gray). Individual task curves for tasks A–D shown in blue, red, green, and purple respectively. See also Figure S1.

measures only for the discrete sample sizes marked on each *x*-axis. Note too that the *x*-axis uses a compressive (square root) scale.

Voxel-level results

Our first analysis assessed the reproducibility of voxel-wise patterns of raw SPM values, which we measured using Pearson correlation of unthresholded P and Q maps. The results of this analysis are shown in Figure 1, which illustrates the results for the average across the four tasks, alongside the average of the hybrid null results across the four tasks. The results using eta^2 rather than Pearson correlation are presented in Figure S1.

There is no universally accepted value for this sort of reproducibility that would allow us to identify a minimum recommended sample size. However, we note that the smallest (measured) sample size for which the average R^2 surpassed 0.5 was 64, which is more than double our standard for a typical sample size.

The results of our second voxel-level analysis, of binary thresholded SPM reproducibility (using Jaccard overlap of maps thresholded using a conservative threshold), are illustrated in Figure 2. Results using a liberal threshold are presented in Figure S2. For these maps, we thresholded to match the proportion of suprathreshold voxels to the observed proportion suprathreshold for each taskÕs thresholded full-sample analysis. That is, differences between tasks in terms of power lead to differences in terms of the proportion suprathreshold, which in turn largely explains the differences between tasks in these four curves. Even at a sample size of 121, the average Jaccard overlap across tasks fails to surpass 0.5.

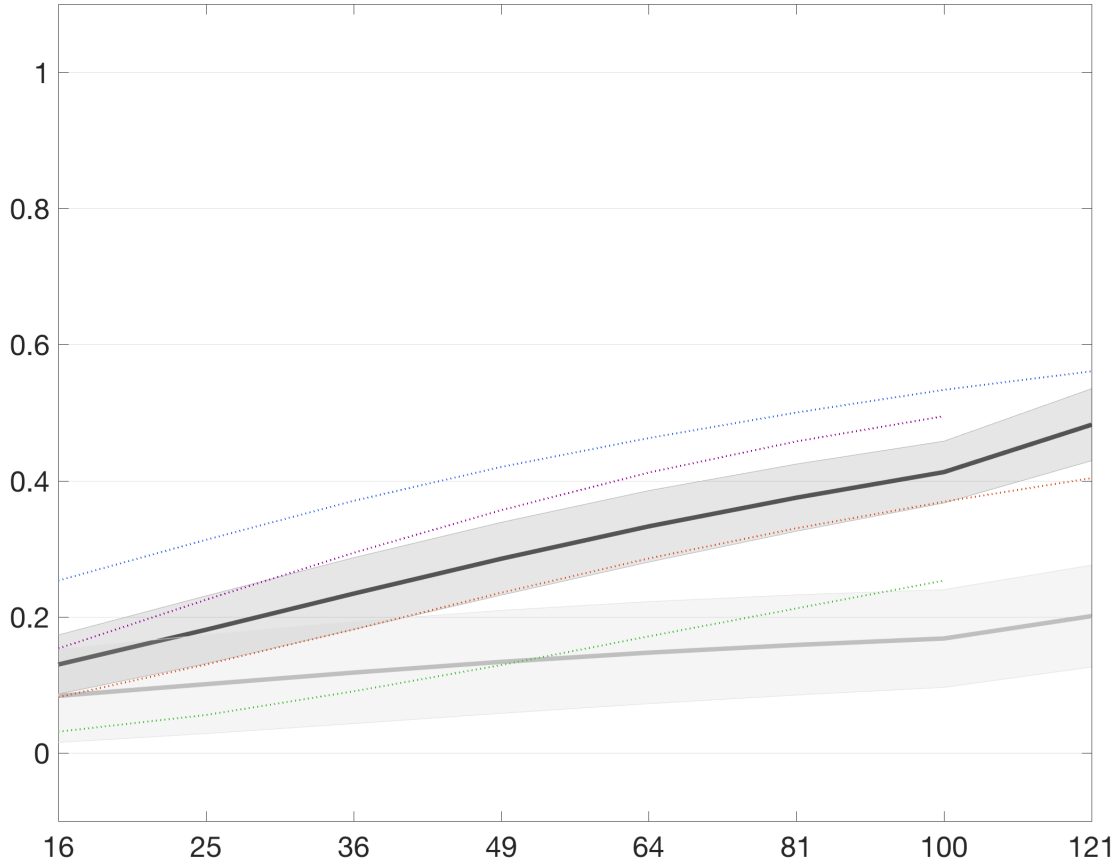


Figure 2. Replicability results for voxel-level (thresholded conservatively) analyses. See Figure 1 for legend. See also Figure S2.

Cluster-level results

The second level at which we considered replicability was at the cluster level. For this analysis, we thresholded each P and Q map using FSL's cluster thresholding tool, and computed the Jaccard overlap between the resulting binarized thresholded maps. Figure 3 presents the results of our cluster-level analyses in terms of mean Jaccard overlap as a function of sample size for each task using the conservative threshold. Results using a liberal threshold are shown in Figure S3. Unsurprisingly, average Jaccard overlap at a sample size of 16 is very near 0, because these SPMs are often null (i.e., contain no suprathreshold voxels), and even when both maps in a pair are non-null, the clusters overlap minimally. However, even at a sample size of 49—the lowest sample size for which fewer than 1% of all SPMs across all tasks are null—the mean overlap is less than 0.3.

Peak-level results

The final level of replicability we considered was at the level of cluster peaks. For this analysis, we assessed how frequently the peak voxel of each cluster was suprathreshold in its corresponding pseudo-replicate. We used a single peak per cluster (i.e., we ignored local maxima), but we did subdivide peaks into those that survived the GRFT-based correction for multiple comparisons, which we refer to as the suprathreshold peaks, and those that did not, which are the subthreshold peaks. The results from these two disjoint sets are shown separately: Figure 4 illustrates the results for suprathreshold peaks, while Figure 5 shows the results for subthreshold peaks, both using a conservative threshold. Results using liberal thresholds are shown in Figures S4–S5.

Note that for Figure 5, because these are peaks that were selected against by our multiple testing correction, lower values are better, in contrast to all the previous plots. High values would potentially reflect overly conservative threshold-

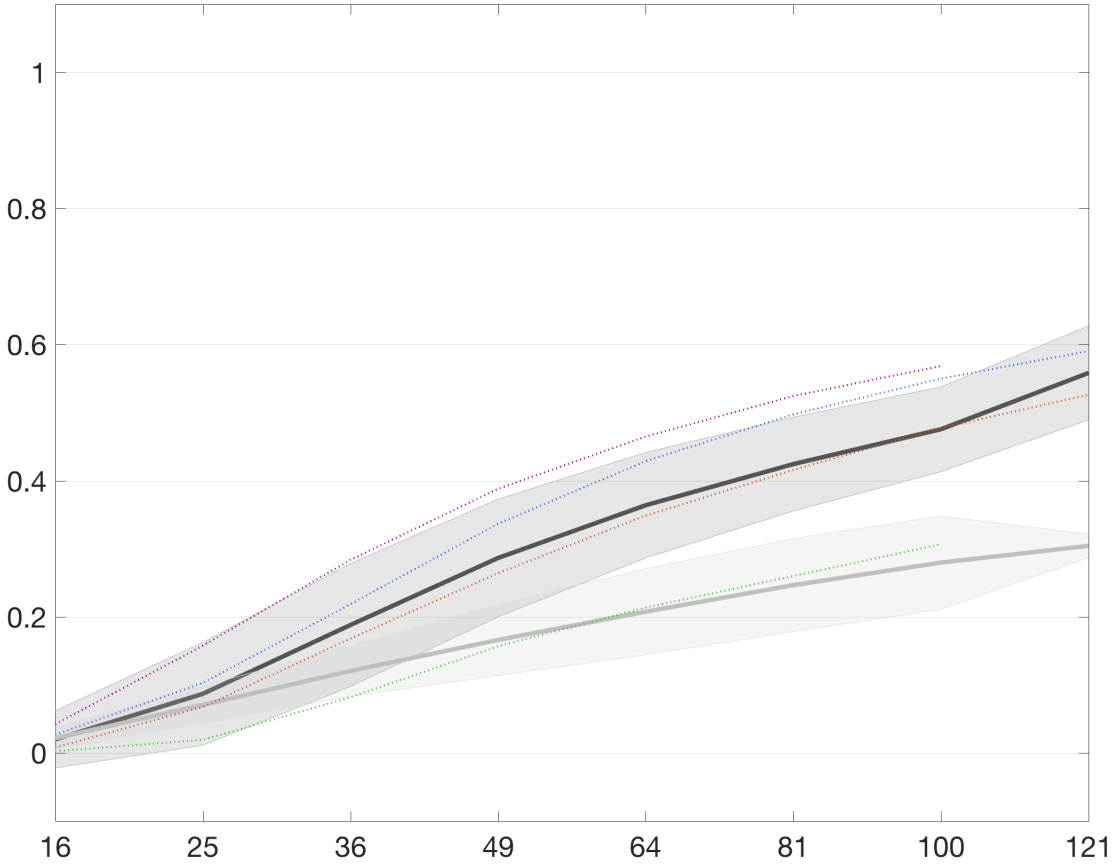


Figure 3. Replicability results for cluster-level (thresholded conservatively) analyses. See Figure 1 for legend. See also Figure S3.

ing, such that many replicable peaks failed to survive.

Measurables results

Our initial modeling suggested that there were at most four variables that contributed to the model fits. In the case of the full model, these variables were: sample size, the average of the mean between-subjects similarity across P and Q, the difference between the mean between-subjects similarity across P and Q, and the average of the mean precision for P and Q. Table 1 presents the standardized beta estimates for each of these variables, along with the unique variance accounted for by each (i.e., above and beyond the other three in a model including only these four variables). Note that because we removed task effects, we reduced the influence of variables that showed differences in range across tasks and exaggerated the influence of variables that did not vary across tasks.

Having estimated *m* to be four, we carried out a double-cross-validation, wherein we assessed how reliably each regressor (from the full set of regressors) was selected as being one of the four to be included in the cross-validation models, as well as how stable each beta estimate was for each regressor. Three variables (sample size, average of the mean between-subjects similarity, and average of the mean precision) were selected at every cross-validation fold, and a fourth variable (average of the standard deviation of between-subjects similarities) was selected in three of four folds (in the fourth fold, the average of the standard deviation of the precision was selected instead). Given the consistency across folds, we present the results for the four most frequently occurring regressors in Table 2.

In addition to these regressor-specific results, we calculated the RMSE of the predicted outcomes for the left-out task for each fold, and compare this against the (biased) standard deviation of the true outcomes, as well as the RMSE of the predicted outcomes from a null model including only

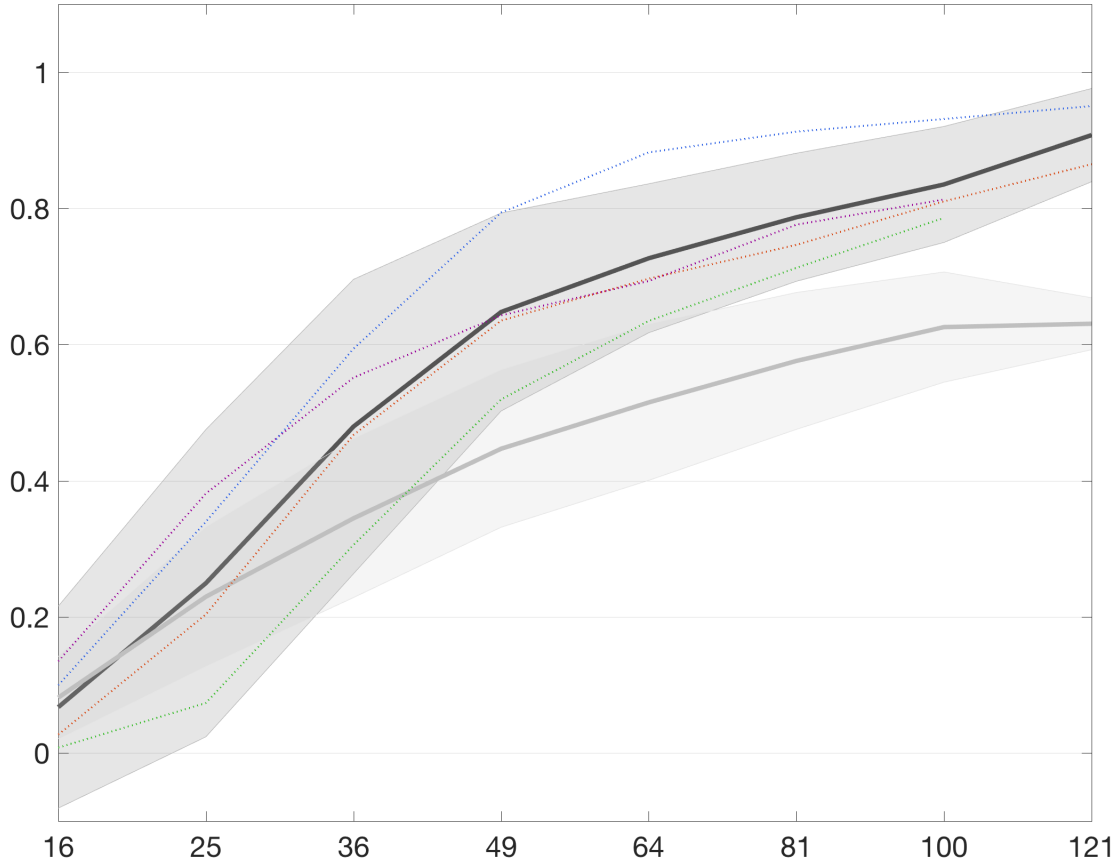


Figure 4. Replicability results for suprathreshold peak-level (thresholded conservatively) analyses. See Figure 1 for legend. See also Figure S4.

Table 1
Results of regression analysis predicting replicability from several variables. Model included only these four variables, and included all four tasks with task effects removed from X and y.

Variable	Standardized beta	$100 \times \Delta R^2$
Sample size	0.751	56.41
Average of mean between-subject similarity	0.100	1.09
Difference of mean between-subject similarity	−0.081	0.71
Average of mean precision	0.070	0.04

an intercept term. The four-variable model RMSE ranged from 0.08–0.233 (mean = 0.141). The (biased) standard deviation ranged from 0.099–0.156 (mean = 0.134). Finally, the intercept-only model RMSE ranged from 0.146–0.270 (mean = 0.213). Note that the biased standard deviation corresponds to an unrealistic model that is allowed to know the mean predicted outcome, so the fact that our four-variable models perform approximately as well is impressive.

Considering both sets of results together—the task-demeaned model with all data (excluding largest sample size results from each task) and the non-demeaned cross-validation models predicting each task based on the other three—it is clear that the two most robust variables are sample size and average of mean between-subject similarity. The latter also has a consistent standardized beta across the two analyses, but the former shows a markedly smaller aver-

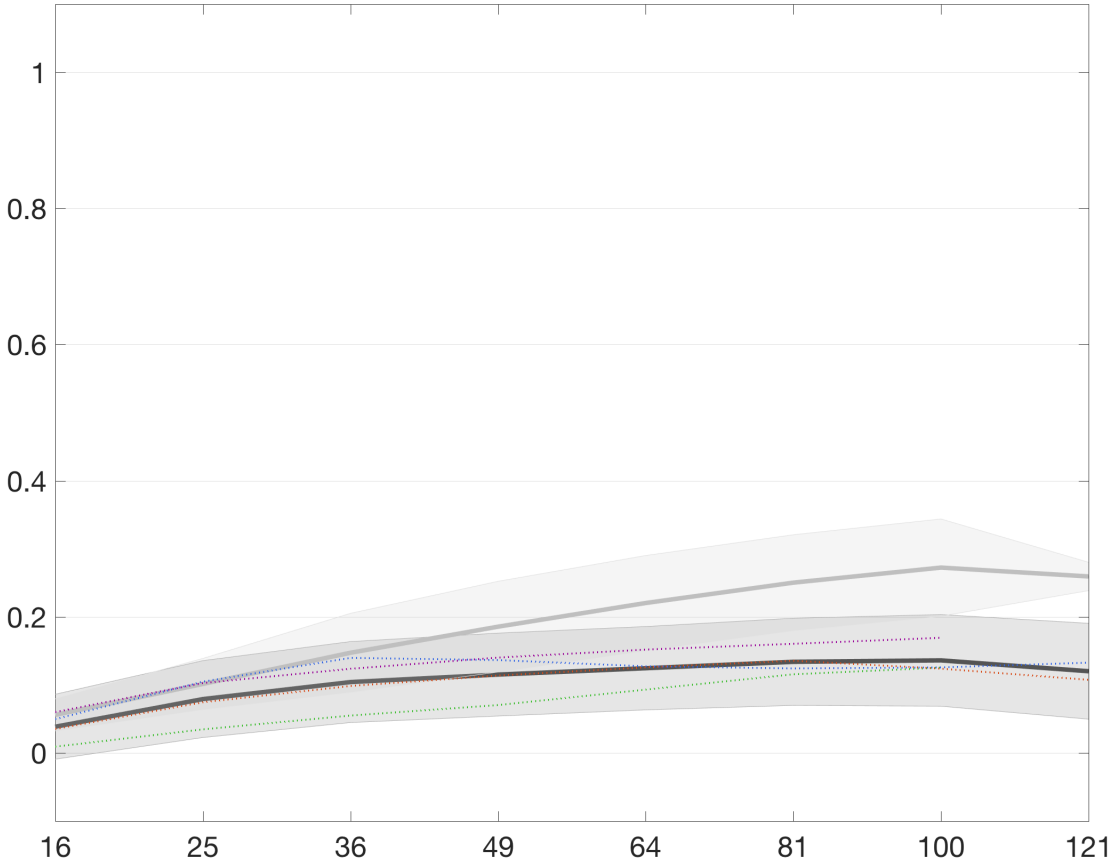


Figure 5. Replicability results for subthreshold peak-level (thresholded conservatively) analyses. See Figure 1 for legend. See also Figure S5.

Table 2
Results of double-CV analysis. The mean beta was computed only over folds for which that variable was included. As a coarser measure of stability, we also present how often, when the beta was included, it had the same sign as the mean.

Variable	Mean standardized beta (when included in model)	# folds sign(beta) matches sign of mean / # folds included
Sample size	0.133	4 / 4
Average of mean between-subject similarity	0.081	4 / 4
Average of SD of between-subject similarity	−0.007	3 / 3
Average of mean precision	+0.000	2 / 4

age standardized beta in the cross-validation analysis. The reason for this can be understood as a result of the differences in demeaning between the two analyses: in the full model, the differences between tasks in terms of mean outcome have been removed, so the curves in Figure 1 will all largely overlap, and sample size will play a relatively consistent role. In the non-demeaned analysis, however, sample size is a much weaker predictor of outcome because the tasks differ so widely due to other factors. In other words, the variability across tasks for a given sample size is a much larger fraction of the variability across sample sizes in the non-demeaned analysis, which translates to lower predictive utility. Nonetheless, the consistency in terms of sign confirms that for these two variables, the same mechanism is at

work at the within- and between-task levels.

Discussion

Despite the development of various tools meant to allow researchers to do prospective power analyses (Mumford & Nichols, 2008; Durnez et al., 2014), such tools are apparently used only infrequently by researchers. Several previous studies have suggested that neuroimaging studies suffer from a marked, possibly fatal, lack of statistical power (Button et al., 2013; Szucs & Ioannidis, 2017). However, Type II errors are not the only problem plaguing neuroimaging, as other studies have demonstrated that certain widely used false-positive correction methods underestimate true false positive rates (Eklund, Nichols, & Knutsson, 2016), and multiple testing correction has been a topic of substantial investigation throughout the history of neuroimaging (Bennett, Miller, & Wolford, 2009).

This previous work has uncovered persistent and troubling problems with standard neuroimaging approaches, particularly as it regards the use of appropriately well-powered (i.e., large) samples. However, no prior work has operationalized reproducibility in the concrete, intuitive ways we have here, nor has any prior work systematically examined the impact of sample size and other dataset properties on such measures of the reproducibility of task-based fMRI.

Our results demonstrate that, regardless of whether one conceptualizes reproducibility as being about patterns at the level of voxels, clusters, or peaks, our estimates of reproducibility at typical sample sizes are startlingly low, particularly when considered in the context of our realistic null results. For instance, the mean between-group Pearson correlation across our four tasks for a sample size of 49 (which is well above the mean or median sample sizes reported in Poldrack et al., 2017 and Szucs & Ioannidis, 2017 over the last several years) is a middling 0.68, compared with an expected mean of 0.56 using our realistic null. Even if one finds our realistic null too stringent, the observed mean tells us that over 50% (53.3%, to be precise) of the variance in voxel intensities between two independent (exact replication) samples of size 49 is unexplained. Such a lack of reproducibility is likely to be concerning to researchers who rely on methods that assume a high degree of spatial specificity (e.g., MVPA). And this same pattern holds true, often to an even greater extent, across our other measures. Furthermore, our results represent a best case scenario for reproducibility (at any of our tested sample sizes for any of our tasks) because we drew samples from the same broad population, we collected all data at one site and one scanner, the experimental methodology and materials were exactly identical for all subjects, and all fMRI data processing was completed using identical processing pipelines on the same computers using the same software. In other words, for any single iteration in our bootstrap method, all pseudo-replicates could be classified as "exact" replications. Deviations from any of these criteria would likely introduce variability in the data collection and processing streams, yielding lower observed reproducibility.

What can explain this pattern of results? Clearly, there are two possible sources of noise in a group-average result: within-subject variance and between-subject variance. Increasing sample size reliably reduces the impact of both sources of noise. However, our analyses of how several easily-measured properties of each data set impacted reproducibility revealed small but consistent roles for several other factors, most notably the mean between-subject similarity, averaged across our pseudo-replicate groups. In fact, between-subject similarity was consistently important in another way as well, as either the difference between the means (in the full model) or the average standard deviation within groups (in our cross-validation analysis) emerged across analyses. Both of these capture the influence of sample heterogeneity.

The idea of inter-individual consistency has been explored previously, and it is not altogether uncommon for researchers to publish maps demonstrating how consistent their results were across participants (e.g., Seghier & Price, 2016). However, our results demonstrate just how substantial an effect individual differences plays. A long line of research has highlighted this extraordinary variability, and argued for taking advantage of this variability, or at least acknowledging and attempting to control for it (Miller et al., 2002; Van Horn, Grafton, & Miller, 2008; Miller et al., 2009, 2012). Our results confirm these earlier observations that individual identity is a powerful driver of patterns of brain activity. Moreover, to the degree that our scanned samples were more homogeneous than the population at large (as is generally the case of scanned samples that largely comprise undergraduates or members of the campus community), it is reasonable to expect that the influence of individual differences would be even larger in any study that used truly representative sampling.

It is possible that our results do a poor job of capturing the average reproducibility that should be expected across the field at large. However, we do not believe this to be the case, for three reasons. First, our results for our tasks A and B, which included an identical set of participants and exactly matched pseudo-replicate groups, span a fairly wide range of reproducibility values. Second, these tasks are well-known (accruing a weighted average of 35.5 citations per year since publication using Google Scholar's citation counts, retrieved 3/21/17), and cover a number of cognitive domains of general interest to researchers in cognitive neuroscience. And third, our results are consistent with earlier work demonstrating the woeful inadequacy of "typical" ($N \simeq 30$) sample sizes. Although there is no simple way to map our results onto these earlier studies, the general conclusion is much the

same.

Although our results clearly point to the insufficiency of typical sample sizes, it would be inappropriate for us to try to use our findings to identify a universal "minimum" sample size that could be adopted across the field. This is because our results do not represent how well sample sizes approximate "ground truth" but rather the expected replicability at each sample size. Moreover, although our tasks cover a reasonable range of effect sizes (as demonstrated by the different reproducibility estimates across tasks), any universal recommendation would have to be made for the smallest "meaningful" effect size, which is not an agreed-upon quantity in the field, and which is probably smaller than the smallest effect size we observed. Instead, we point readers to existing tools for conducting prospective power analyses, and hope that future research will develop similar tools that make use of the replicability measures we have employed here.

Our hope is that whereas earlier work pointing out the ubiquity of underpowered studies may have been seen by the average researcher as too abstract or technical to worry about, the present results are straightforward enough that researchers have no choice but to confront the fact that "typical" sample sizes cannot be trusted to produce reproducible results, irrespective of how reproducibility is measured. Thus, our results add to the growing consensus calling for a paradigm shift in the field, away from small-scale studies of hyper-specific processes to large-scale studies designed to address multiple theoretical questions at once. Alternatively, methods which are transparent about treating individuals as unique—for instance, individual differences approaches (Van Horn et al., 2008) or encoding methods (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009)—likely deserve more attention for their potential to overcome at least one part of the problem with small samples (i.e., individual variability).

Conclusion

Reproducibility is the foundation of scientific progress. Unfortunately, for a variety of reasons, many scientific fields are currently gripped by a crisis of irreproducibility (Baker et al., 2016). While some of the causes of this crisis are deeply interwoven into the academic landscape—incentives related to publication, funding, and tenure—the most straightforward solution relates to statistical power. Researchers in fMRI may have believed that they were adequately addressing concerns about power by using carefully optimized designs and rule-of-thumb "large enough" sample sizes (Friston, 2012; Liu et al., 2001). Indeed, the success of quantitative meta-analysis methods (e.g., activation likelihood estimation; Eickhoff, Bzdok, Laird, Kurth, & Fox, 2012), alongside reports of moderate test-retest reliability for task-based fMRI (Bennett & Miller, 2010), may have reinforced the sense that power in task-based fMRI was a solved problem. However, meta-analytic approaches work precisely by relaxing specificity about spatial location (and in many cases, about design features including task, contrast, or putative cognitive processes); likewise, test-retest reliability is only weakly related to reproducibility. Despite empirical work demonstrating that typical fMRI sample sizes are inadequate, there seems to be little motivation to change the status quo (Button et al., 2013; Szucs & Ioannidis, 2017). Our results unambiguously demonstrate that reproducibility (as measured at multiple levels of analysis) is strikingly low at "typical" sample sizes, thus serving to highlight and extend these previous results. The solution to this problem may be arduous for researchers and funding agencies, for instance requiring a paradigm shift away from incremental research using bespoke experiments with small samples. However, if our goal is the advancement of scientific understanding, the status quo—thousands of underpowered and minimally-reproducible papers published annually—clearly cannot continue.

Acknowledgments

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract 2014-13121700004 to the University of Illinois at Urbana-Champaign (PI: Barbey). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

The authors would like to thank Soohyun Cho, Keith Holyoak, Michael Stevens, Jeremy Gray, Todd Braver, and Debbie Hannula for graciously providing original task design, timing, and stimulus files for the tasks reported in this manuscript.

References

Andersson, J. L., Jenkinson, M., Smith, S., et al. (2007). Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, *2*.

Baker, M., et al. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452–454.

Barnes, R. M., Tobin, S. J., Johnston, H. M., MacKenzie, N., & Taglang, C. M. (2016). Replication rate, framing, and format affect attitudes and decisions about science claims. *Frontiers in Psychology*, *7*.

Bennett, C. M., Miller, M., & Wolford, G. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for multiple comparisons correction. *Neuroimage*, *47*(Suppl 1), S125.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*(1), 133–155.

Bennett, C. M., & Miller, M. B. (2013). fmri reliability: influences of task and experimental design. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 690–702.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, *63*(1), 289–300.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., ... Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral cortex*, *20*(3), 524–533.

Durnez, J., Degryse, J., Moerkerke, B., Seurinck, R., Sochat, V., Poldrack, R., & Nichols, T. (2016). Power and sample size calculations for fmri studies based on the prevalence of active peaks. *bioRxiv*, 049429.

Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fmri. *NeuroImage*, *84*, 45–64.

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *Neuroimage*, *59*(3), 2349–2361.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413.

Evans, S. (2017). What has replication ever done for us? insights from neuroimaging of speech perception. *Frontiers in human neuroscience*, *11*.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *Neuroimage*, *61*(4), 1300–1310.

Gabrieli, J. D., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, *85*(1), 11–26.

Gonzalez-Castillo, J., & Talavage, T. M. (2011). Reproducibility of fmri activations associated with auditory sentence comprehension. *Neuroimage*, *54*(3), 2138–2155.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience*, *6*(3), 316–322.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63–72.

Hannula, D. E., & Ranganath, C. (2008). Medial temporal lobe activity predicts successful relational memory binding. *Journal of Neuroscience*, *28*(1), 116–124.

Ingre, M. (2013). Why small low-powered studies are worse than large high-powered studies and how to protect against Òtrivial Ó findings in research: Comment on friston (2012). *Neuroimage*, *81*, 496–498.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, *5*(2), 143–156.

Liu, T. T., & Frank, L. R. (2004). Efficiency, power, and entropy in event-related fmri with multiple trial types: Part i: Theory. *NeuroImage*, *21*(1), 387–400.

Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fmri. *Neuroimage*, *13*(4), 759–773.

Miller, M. B., Donovan, C.-L., Bennett, C. M., Aminoff, E. M., & Mayer, R. E. (2012). Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *Neuroimage*, *59*(1), 83–93.

Miller, M. B., Donovan, C.-L., Van Horn, J. D., German, E., Sokol-Hessner, P., & Wolford, G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *Neuroimage*, *48*(3), 625–635.

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., ... Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*, *14*(8), 1200–1214.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fmri studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, *39*(1), 261–268.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., ... others (2012). Test–retest reliability of evoked bold signals from a cognitive–emotive fmri test battery. *Neuroimage*, *60*(3), 1746–1758.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*.

Replication studies offer much more than technical details. (2017, Jan. 19). *Nature*, *541*, 259–260.

Seghier, M. L., & Price, C. J. (2016). Visualising inter-subject variability in fmri using threshold-weighted overlap maps. *Scientific reports*, *6*.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17*(3), 143–155.

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing n. *Frontiers in psychology*, *7*.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neu-

roscience and psychology literature. *PLoS biology*, *15*(3), e2000797.

Turner, B. O., & Miller, M. B. (2013). Number of events and reliability in fmri. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 615–626.

Van Horn, J. D., Grafton, S. T., & Miller, M. B. (2008). Individual variability in brain activity: a nuisance or an opportunity? *Brain imaging and behavior*, *2*(4), 327.

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fmri: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293–309.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*.

Witt, S. T., & Stevens, M. C. (2013). fmri task parameters influence hemodynamic activity in regions implicated in mental set switching. *NeuroImage*, *65*, 139–151.

Woolrich, M. (2008). Robust group analysis using outlier inference. *Neuroimage*, *41*(2), 286–301.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *NeuroImage*, *23*, S189–S195.

Zhang, H., Nichols, T. E., & Johnson, T. D. (2009). Cluster mass inference via random field theory. *Neuroimage*, *44*(1), 51–61.

Supplemental Materials

Task descriptions (design and analysis)

The design of each task was based closely on a previously-published instantiation of each task. Here, we provide the basic details of each task, and explicitly highlight any points at which the design or analysis deviated from its previously-published antecedent.

Task A. See Cho et al. (2010) for full details regarding the paradigm. This was a task of analogical reasoning, with a 2×2 design in which relational complexity (the number of to-be-attended stimulus traits, 1 or 3) was crossed factorially with interference level (the number of irrelevant dimensions that lead to an incorrect response, 0 or 1). In our adaptation of their design, we included three functional runs, each of which contained 54 trials. These trials were modeled by seven (RT-duration) regressors: four defined per the 2×2 design described above; another two for invalid trials (relational complexity 1 or 3); and a final regressor for error trials. Our primary contrast of interest compared relational complexity 1 with relational complexity 3, collapsing across interference levels. On average per run, this contrast included 18.5 trials (standard deviation across participants $=$ 1.3 trials) versus 17.1 trials (standard deviation $=$ 2.4 trials).

Task B. See Witt and Stevens (2013) for full details regarding the paradigm. This was a task of set switching. Participants were always tasked with counting the number of unique levels of a given relevant dimension; the relevant dimension changed (as indicated by a printed cue above the stimulus) every 1–6 trials. Trials varied in terms of: switch vs. non-switch (as well as number of preceding non-switch trials for switch trials); stimulus complexity (1, 2, or 3 varying dimensions with multiple levels); and response complexity (1, 2, or 3 potential valid response options across all dimensions). As in Witt and Stevens (2013), there were two functional runs, each with 81 trials. These trials were modeled with ten (RT-duration) regressors: two for switch/non-switch; six parametric regressors (orthogonalized with respect to the switch/non-switch EVs) encoding separately for switch and non-switch trials stimulus complexity, response complexity, and number of preceding non-switch trials; and two regressors to model error and post-error trials. Our primary contrast of interest compared switch and non-switch trials. On average per run, this contrast included 31.0 trials (standard deviation $=$ 5.6 trials) versus 32.7 trials (standard deviation $=$ 5.0 trials).

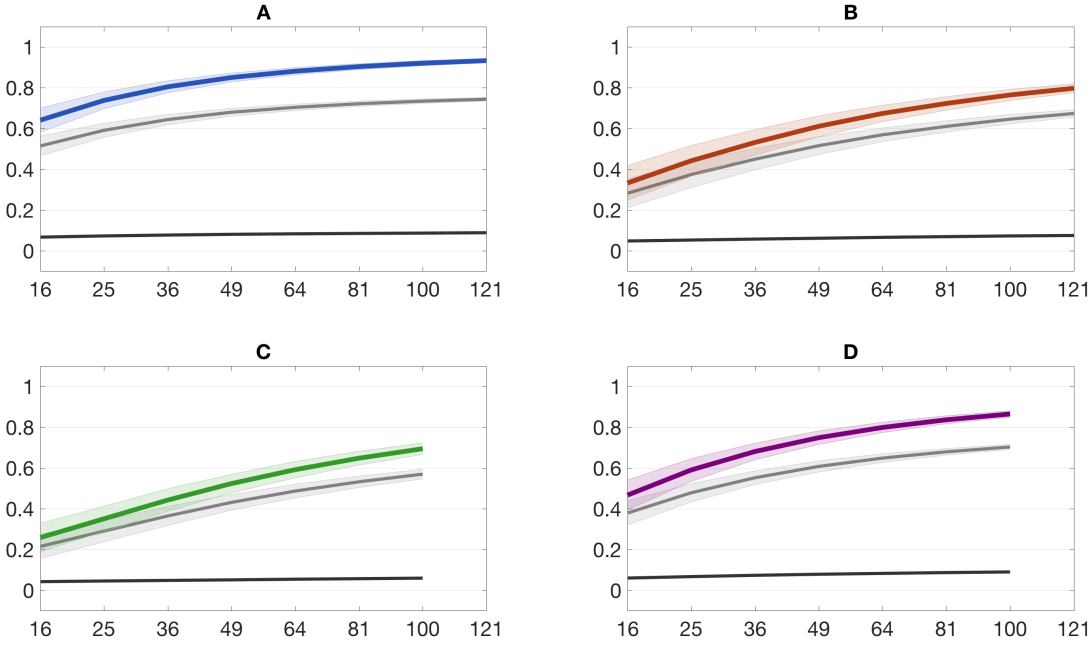
Task C. See Gray, Chabris, and Braver (2003) for full details regarding the paradigm. This was a 3-back working memory task. Participants saw multiple short series of consecutive stimuli, during which they had to respond to items that had appeared exactly three items earlier ("targets"). These were intermixed with new items, as well as items that had appeared either two, four, or five items earlier ("lures"). As in Gray et al. (2003), there were two functional runs (one using faces, the other using words, order counterbalanced across participants), each of which included four blocks of 16 trials (plus five jitter fixation trials per block). Trials were modeled with seven regressors: two each (correct/incorrect) for targets, lures, and non-lures; and one for missed trials. Our primary contrast of interest compared correct targets and correct lures. On average per run, this contrast included 10.1 trials (standard deviation $=$ 2.7 trials) versus 12.8 trials (standard deviation $=$ 2.3 trials).

Task D. See Hannula and Ranganath (2008) for full details regarding the paradigm. This was a task of relational memory. Participants viewed displays of four 3D objects on a 3×3 grid, and had to indicate whether a test grid, displayed rotated after a short delay, matched the original layout. These test grids could be of three types: "match," in which all items retained their original relative positions; "mismatch," in which one item moved out of position; or "swap," in which two items swapped positions. Each trial was comprised of an encoding period, a delay period, and a test period. There were five functional runs, each of which included 15 trials. These trials were modeled with a simplified set of four regressors: one each for correct encoding+delay periods (collapsed across trial types), match test periods, and non-match test periods (collapsing across "mismatch" and "swap" trials); and one for all periods of all incorrect trials. Our primary contrast of interest compared correct match and non-match test periods. On average per run, this contrast included 3.9 trials (standard deviation $=$ 0.7 trials) versus 5.7 trials (standard deviation $=$ 1.9 trials).

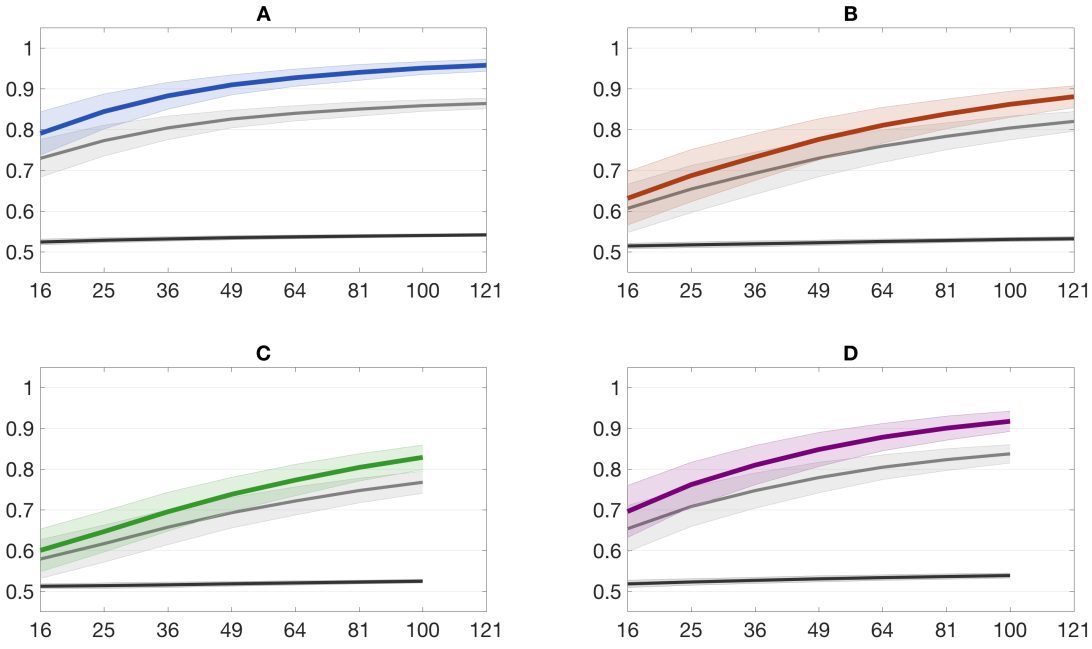
Peak height reproducibility analysis

For our peak analyses, in contrast to our other analysis approaches, it is possible to construct a disaggregated statistic—that is, to define reproducibility on a peak-by-peak basis, rather than only mapwise. This allows us to look in a more fine-grained manner at the relationship between effect size, reproducibility, and sample size. To this end, we collated each peak z value with whether that voxel was reproduced (i.e., was suprathreshold in the counterpart map), separately for each sample size but combining across all tasks. We then fit a separate logistic regression model for each sample size. We then used the `sim` function (part of the `arm` package in **R**) to graphically display uncertainty around the model fits, which is especially pronounced for values of peak z outside of the range observed for a given sample size. Note that this approach treats task as a fixed effect, and moreover, weights tasks proportional to the number of total peaks across all maps for a given sample size. Note too that, as with the main peak analysis reported in the manuscript, a low $p(\text{reproduced})$ is heavily influenced by the sparsity of the counterpart map. The results of this analysis are shown in Figure S6.

Supplemental figures

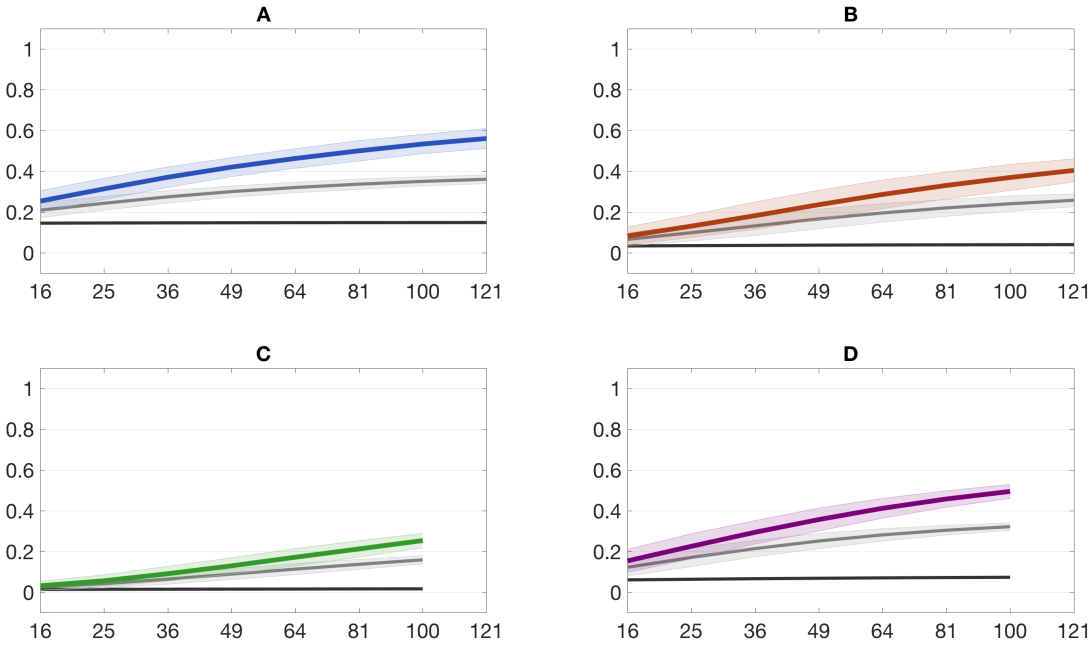


(a)

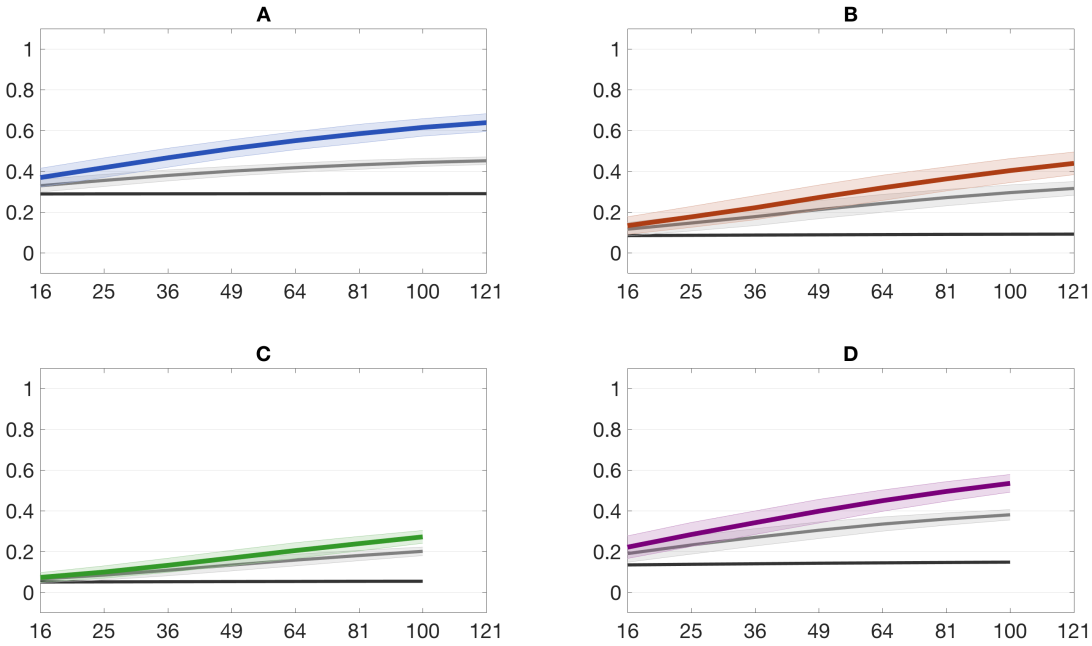


(b)

Figure S1. Replicability results for voxel-level (unthresholded) analyses. Panes show results for tasks A–D. Each pane presents the observed replicability for that task in the same color as used in Figure 1 (± 1 standard deviation), along with the corresponding strong (in light gray) and weak (in dark gray) null results. (a) Correlation; (b) eta^2.

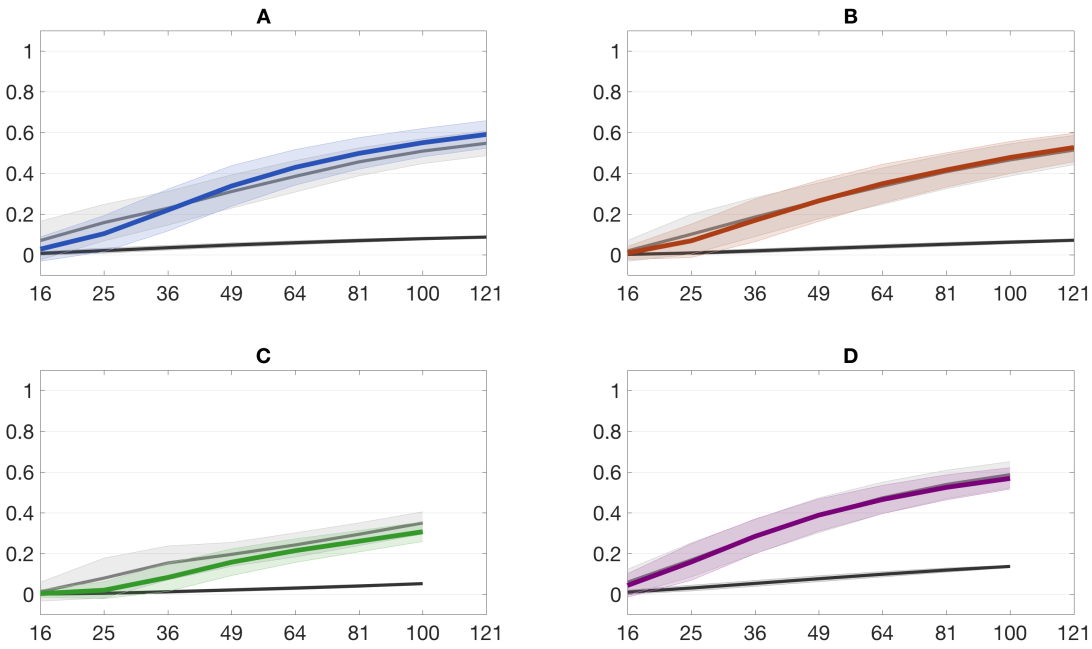


(a)

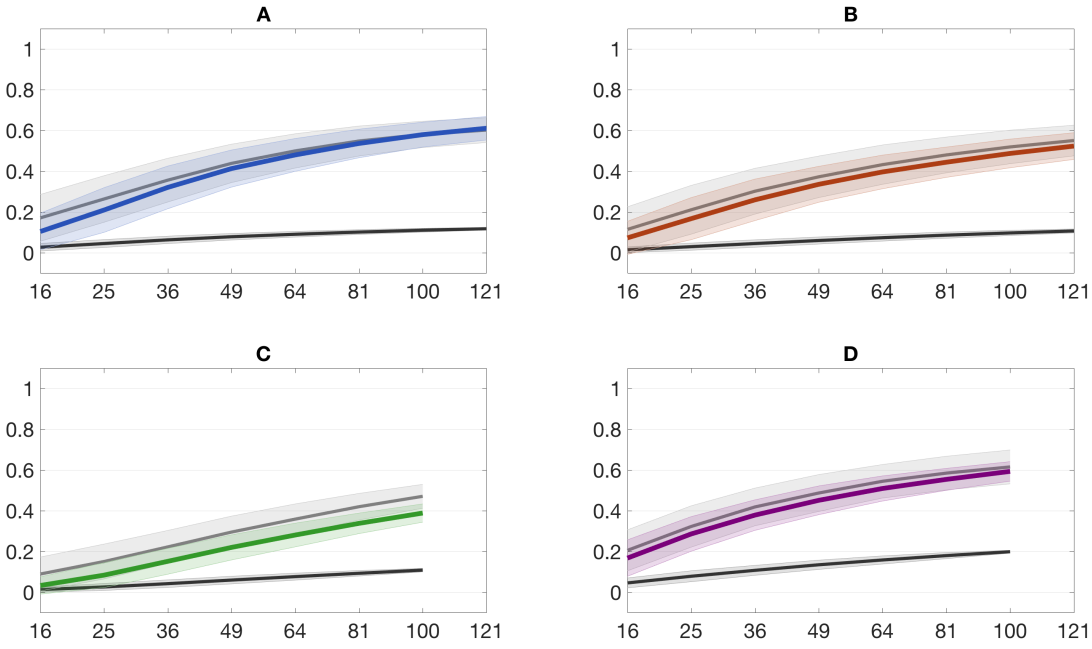


(b)

Figure S2. Replicability results for voxel-level analyses. See Figure S1 for legend. (a) Conservative threshold; (b) Liberal threshold.

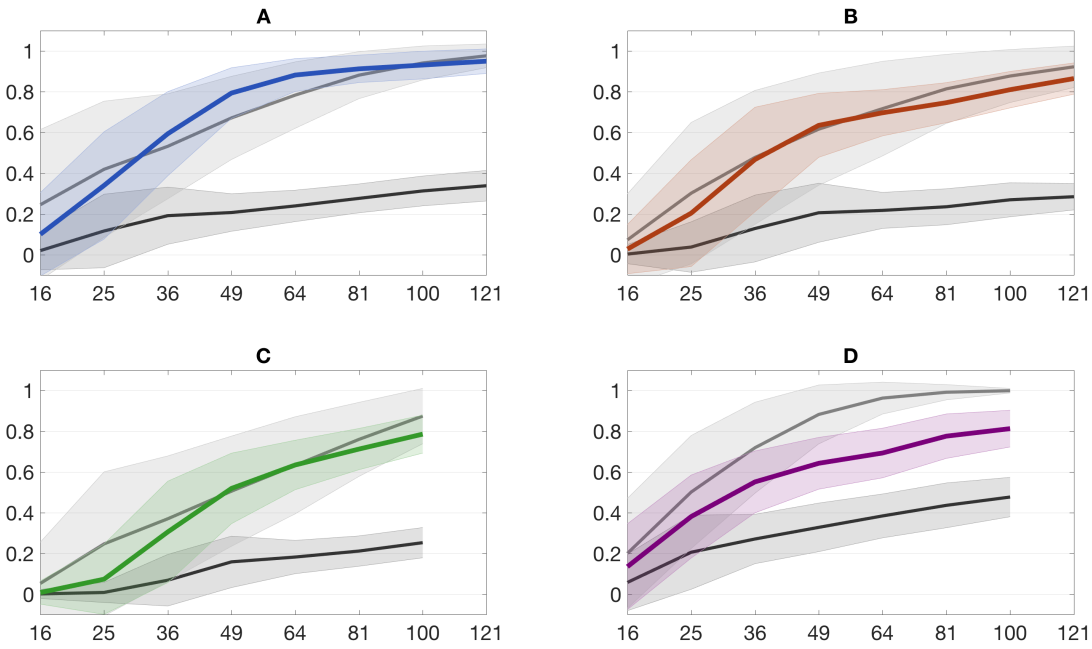


(a)

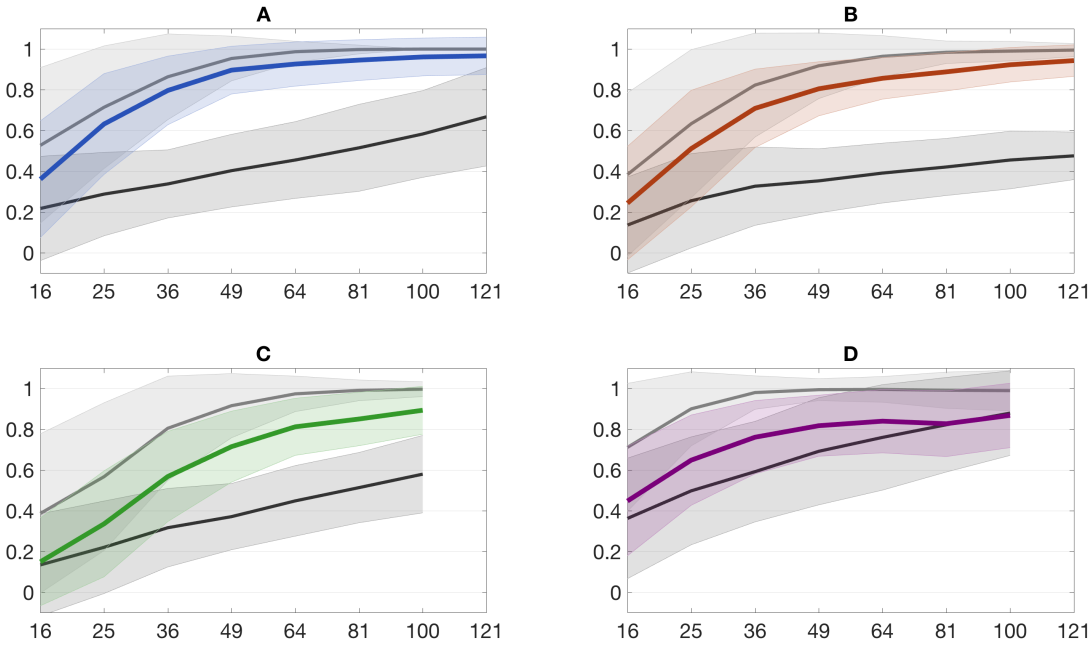


(b)

Figure S3. Replicability results for cluster-level analyses. See Figure S1 for legend. (a) Conservative threshold; (b) Liberal threshold.

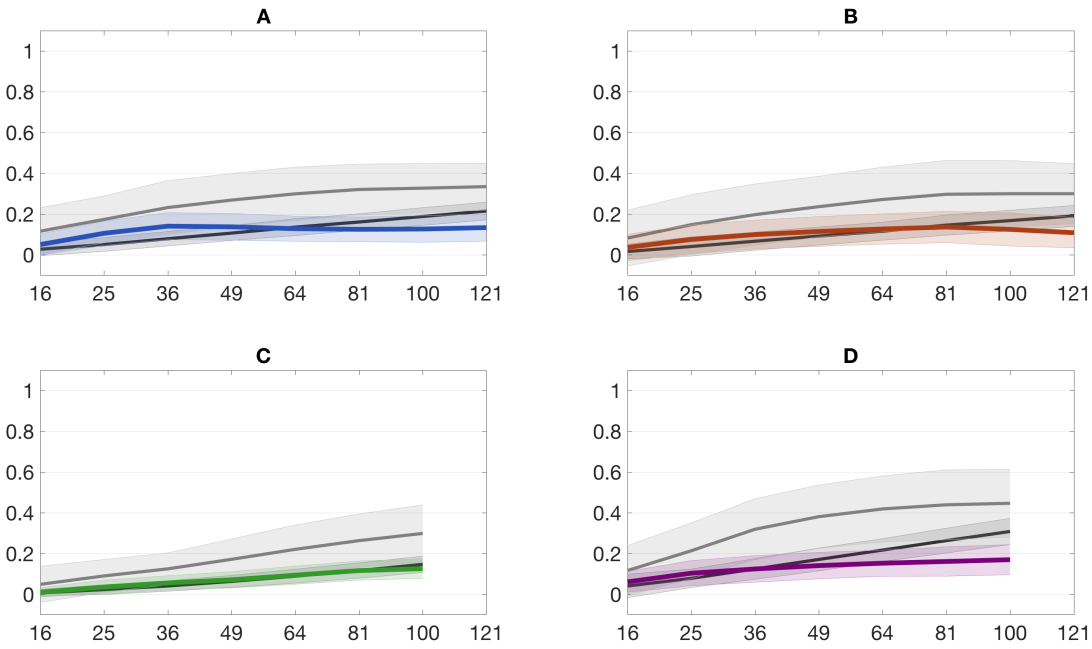


(a)

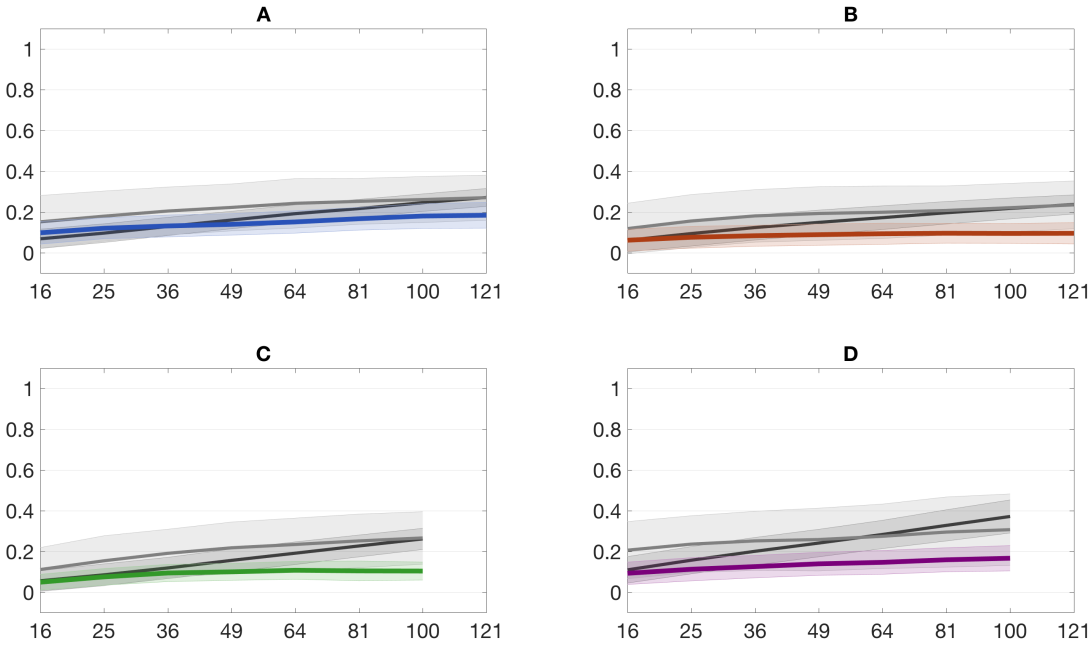


(b)

Figure S4. Replicability results for suprathreshold peak-level analyses. See Figure S1 for legend. (a) Conservative threshold; (b) Liberal threshold.

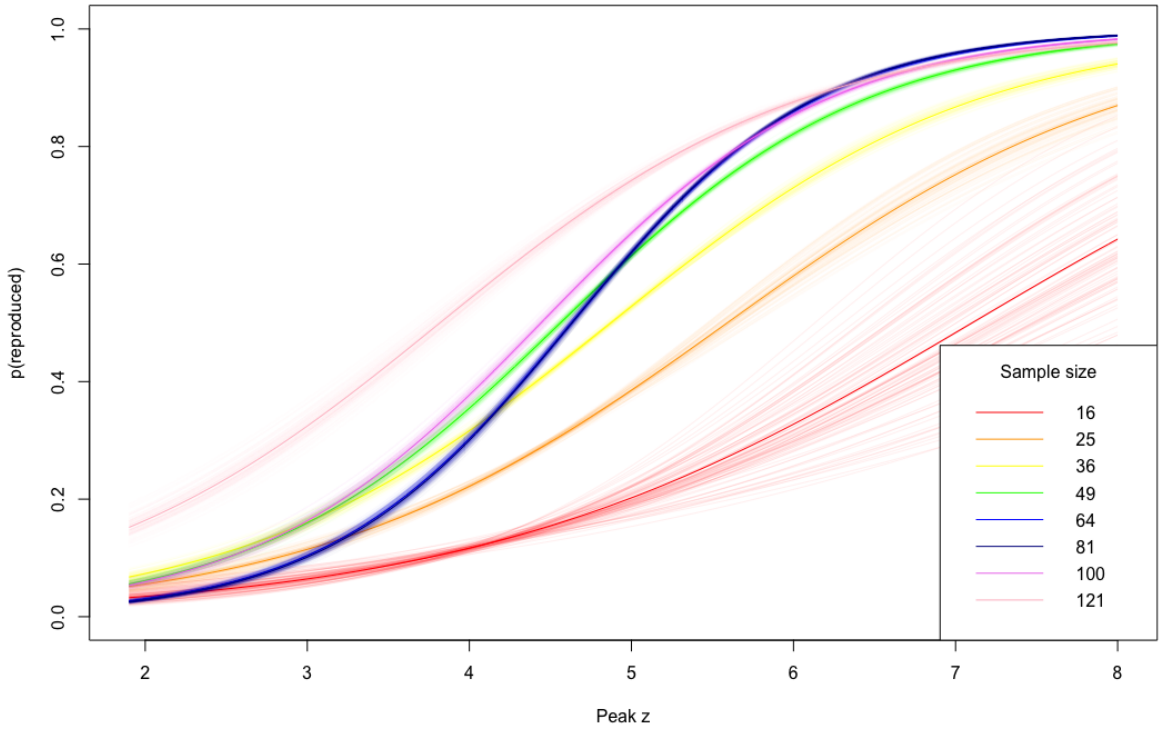


(a)

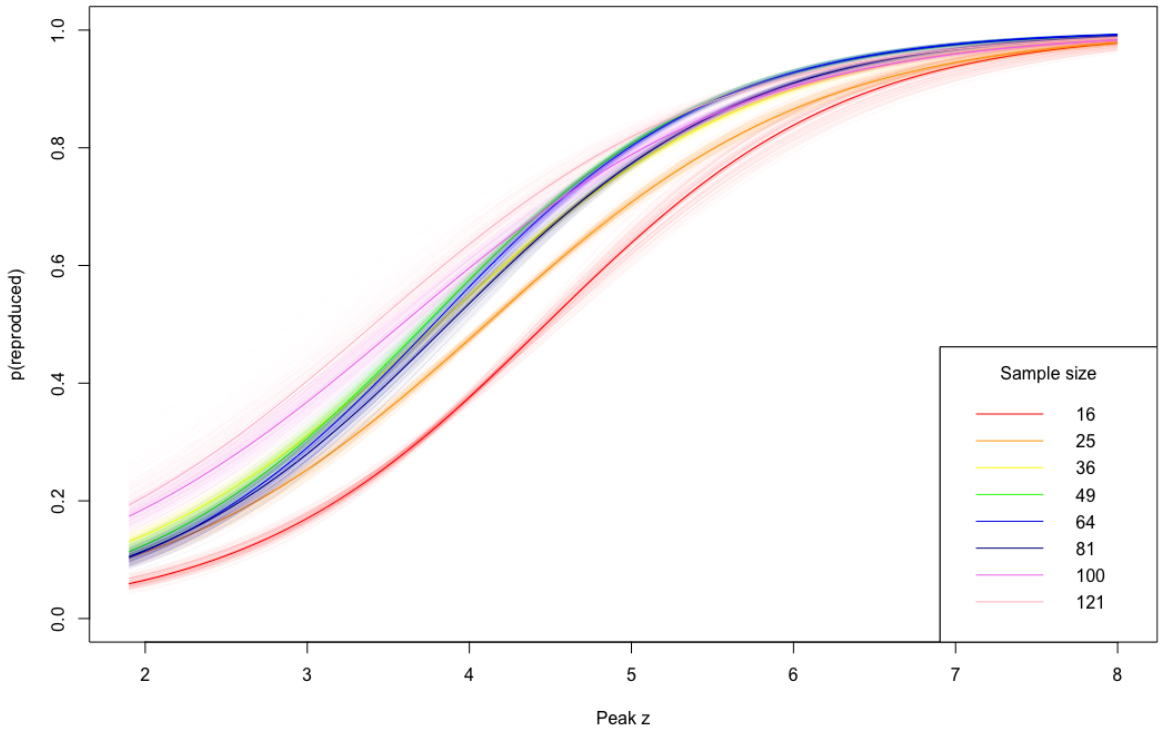


(b)

Figure S5. Replicability results for subthreshold peak-level analyses. See Figure S1 for legend. (a) Conservative threshold; (b) Liberal threshold.



(a)



(b)

Figure S6. Replicability as a function of peak *z* value and sample size. (a) Conservative threshold; (b) Liberal threshold.