COMMUNICATIONS BIOLOGY

ARTICLE

DOI: 10.1038/s42003-018-0073-z OPEN

Small sample sizes reduce the replicability of task-based fMRI studies

Benjamin O. Turner[1], Erick J. Paul[2], Michael B. Miller[3] & Aron K. Barbey[4,5,6,7,8,9]

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 eleven tasks, we demonstrate the impact of sample size on replicability, assessed at different levels of analysis relevant to fMRI researchers. We find that the degree of replicability for typical sample sizes is modest and that sample sizes much larger than typical (e.g., $N = 100$) produce results that fall well short of perfectly replicable. 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 replicability, our hope is that our results are more understandable and convincing to researchers who may have found previous results advocating for larger samples inaccessible.

[1] Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore 639798, Singapore. [2] Microsoft Corporation, 1 Microsoft Way, Redmond, WA 98052, USA. [3] Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA 93106, USA. [4] Department of Psychology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [5] Neuroscience Program, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [6] Department of Bioengineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [7] Center for Brain Plasticity, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [8] Carle R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. [9] Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. These authors contributed equally: Benjamin O. Turner, Erick J. Paul Correspondence and requests for materials should be addressed to A.K.B. (email: barbey@illinois.edu)

Recent years have seen an increased focus on what has been referred to as a reproducibility crisis in science, both in science at large[1–3], and perhaps even more acutely in the psychological sciences[4]. Some of the reasons behind this crisis—including flawed statistical procedures, career incentive structures that emphasize rapid production of splashy results while punishing studies that report null findings, and biases inherent in the publication system—have been articulated carefully in previous work, again both generally[5–7], and for fMRI in particular[8–13]. 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., ref. [14]; cf. ref. [15]), 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, or is based on outdated procedures[12].

Of course, this lack of power is driven in large part by the great expense associated with collecting fMRI data[16]. 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[16,17], researchers may struggle to understand these tools, because defining power in an fMRI context involving tens or even 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 generally fair-to-good test-retest reliability[18–21]. It is possible that researchers take this to mean that large samples are not necessary, particularly if the researcher misunderstands standard design optimization approaches for increasing power at the individual level to mean their small samples are sufficiently powered[22–24]. 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, 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-replicable 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[25–27].

Our goal in the present study is to provide empirical estimates of fMRI's replicability—approximated using a resampling-based pseudo-replication approach within a pair of large datasets—in terms of the levels of results that are useful in the field (i.e., multi-voxel patterns of unthresholded activity 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 might affect replicability, including design power[28]. We present the result from each of the three levels of analysis described in the Methods—voxel, cluster, and peak—in separate sections below. For all Figures throughout the first three sections, note that we plot results as lines for clarity, but computed our measures only for the discrete sample sizes marked on each *x*-axis. Note too that the *x*-axis uses a compressive (square root) scale. 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. In a separate section, we explore the relationship between various measurable properties of the data and the voxel-level replicability results.

We emphasize that our results, far from being relevant only to researchers whose specific interest is in studying reproducibility or replicability (e.g., ref. [29]), are applicable to all researchers who are interested in using fMRI to produce valid and meaningful neuroscientific discoveries. In fact, we use $N \approx 30$ as our standard for a typical (i.e., normative relative to the prevailing standards) fMRI sample size, which is in line with empirical estimates by[10] (median sample size of fMRI studies in 2015 = 28.5) and[11] (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 refs. [9] and ref. [11]: replicability at typical sample sizes is relatively modest. Furthermore, sample size is the largest driver of replicability among those that we examined. Considering at least some of our measures, researchers may wish to consider alternative approaches to answering their questions, rather than facing disappointingly low replicability even at large (and expensive) sample sizes.

Results

Shared analysis setup. All of the results described below are based on a shared initial step, in which we resample participants without replacement to produce pairs of disjoint groups with a matched sample size. For any given resampling, we arbitrarily label these two maps the "P" and "Q" maps.

Voxel-level results. Our first analysis assessed the replicability of voxelwise patterns of raw Statistical Parametric Map values (SPM; in this case representing parameter estimates from the general linear model and not to be confused with the software package of the same name). For this analysis, we measured replicability using a Pearson correlation between vectorized unthresholded P and Q maps. The results of this analysis are shown in Fig. 1, which illustrates the results for the average across the eleven tasks, alongside the average of the null results across the tasks.

There is no universally accepted value for this sort of replicability 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 (the point at which more variance between the paired maps is explained than unexplained) was 36, which is still larger than our standard for a typical sample size.

The results of our second voxel-level analysis, of binary thresholded SPM replicability (using Jaccard overlap of maps thresholded proportionally using a conservative threshold), are illustrated in Fig. 2. Results using a liberal threshold are presented in Supplementary Fig. 1. For these maps, we thresholded to match the proportion of suprathreshold voxels to the observed proportion suprathreshold for each task's 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 eleven curves. Even at a sample size of 121, the average Jaccard overlap across tasks fails to surpass 0.6.

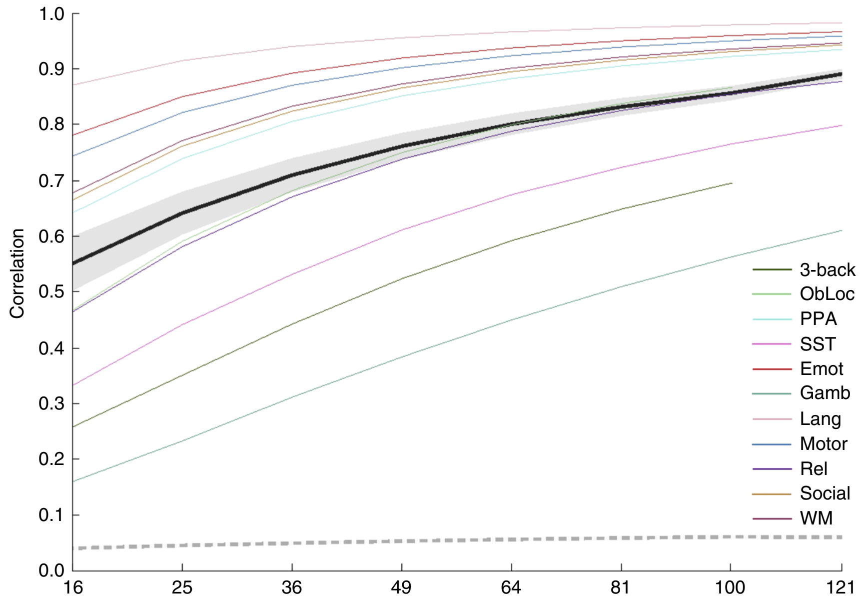

Fig. 1 Unthresholded voxel-level results. Replicability results for voxel-level (unthresholded) analyses, measured as the Pearson correlation between paired group maps. Average observed (±1 mean within-task standard deviation) shown in black (dark gray); average null (±1 standard deviation) shown in dashed medium gray (light gray). Also shown are individual task curves for each task (colors given in legend; refer to Table 1 for task abbreviations)

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 the cluster thresholding tool from the analysis suite FSL, 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 Supplementary Fig. 2. Unsurprisingly, average Jaccard overlap at a sample size of 16 is near 0 for several tasks, 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. As with the analyses holding a set proportion suprathreshold per task, mean overlap remains below 0.5 (the point at which more voxels are overlapping than non-overlapping) up to a sample size of at least 81.

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). Figure 4 illustrates the results for suprathreshold peaks. Results using a liberal threshold are shown in Supplementary Fig. 3. On average across tasks, even with a sample size of 121, the peak voxel failed to surpass threshold in its corresponding pseudoreplicate over 20% of the time.

Results of measureables analyses. Although our focus was on the effect of sample size on replicability, and this variable was the only one we manipulated systematically in our pseudoreplicate analysis, we nonetheless have access to a number of other variables whose potential influence on replicability we can measure, including variables related to motion, contrast power, and within- and between-individual variability. Our goal in these measurables analyses is not to draw strong inferences about the influence of each of these variables, which is complicated by the interdependence between many of our observations (e.g., overlap between pseudoreplicates in terms of which participants comprise each group; overlap between sample sizes for the same pseudoreplicate, with larger sample sizes constituting supersets of smaller sample sizes; and overlap between tasks in terms of participant identity). Instead, we present qualitative results of a simple analysis designed to provide some intuitive understanding of the relative role each variable plays in driving replicability. Full details on the modeling approach can be found in the Methods, but briefly, for each task we fit a separate simple regression model relating each of the above-mentioned variables to replicability, and take as our measures of interest the effect size and ΔR^2 associated with each variable.

Only two variables emerged as qualitatively having more than a very weak relationship with replicability. As expected, sample size was related to replicability ($d_{\mathrm{MAP}} = 1.71$, $\Delta R^2_{\mathrm{MAP}} = 0.37$). The second variable to demonstrate a relationship with replicability was between-individual variability ($d_{\mathrm{MAP}} = 0.27$, $\Delta R^2_{\mathrm{MAP}} = 0.01$). Although several other variables demonstrated a consistent relationship with replicability across tasks, the effects were minuscule. Results for all variables are given in Supplementary Table 1.

Discussion

Despite the development of various tools meant to allow researchers to do prospective power analyses[15,16], 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[9,11]. 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 actual false positive rates[30], and multiple testing correction has been a topic of substantial investigation throughout the history of neuroimaging[31].

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, there have been few previous attempts to operationalize replicability in the concrete ways we have here, or to systematically examine the impact of sample size and other dataset properties on such measures of replicability. One exception is[12], which is a spiritual predecessor to the current work.

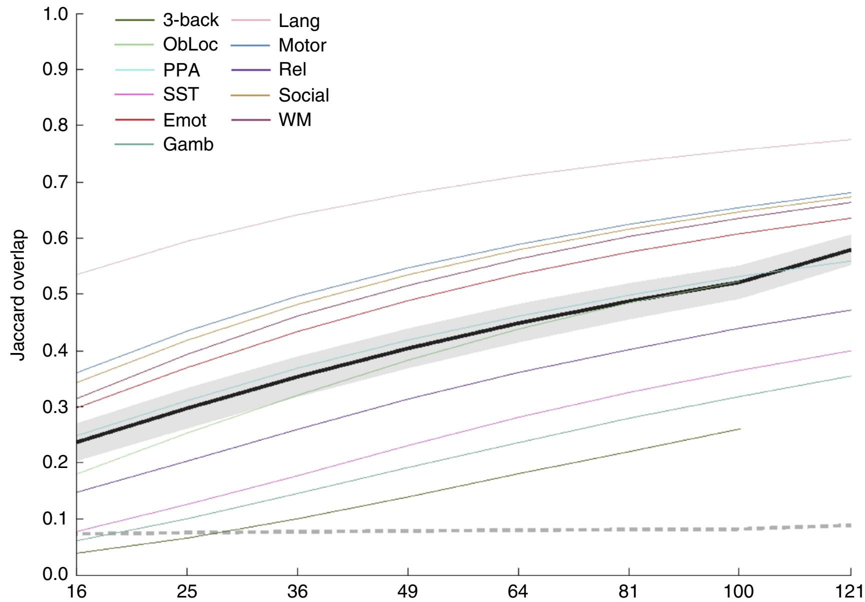

Fig. 2 Conservatively thresholded voxel-level results. Replicability results for voxel-level (thresholded conservatively) analyses, measured as the Jaccard overlap between paired group maps—that is the ratio of the number of voxels in the intersection of the two thresholded maps to the number of voxels in the union of the two. Average observed (±1 mean within-task standard deviation) shown in black (dark gray); average null (±1 standard deviation) shown in dashed medium gray (light gray). Also shown are individual task curves for each task (colors given in legend; refer to Table 1 for task abbreviations). See also Supplementary Fig. 1

However, in addition to being a decade old (during which time the field of fMRI research has changed substantially), the measures used by[12] may not be as intuitively accessible as those we have adopted here. Furthermore, that work focused on many other drivers of replicability (e.g., threshold choice, thresholding method), and their main conclusion with respect to sample size is that $N = 20$ should be considered a minimum—a value that our results suggest is too low.

Our results demonstrate that, regardless of whether one conceptualizes replicability as being about patterns at the level of voxels, clusters, or peaks, our estimates of replicability at typical sample sizes are quite modest. For instance, the mean peak hit rate across our eleven tasks for a sample size of 36 (which is above the mean or median sample sizes reported in[10,11] over the last several years) is below 0.5. The observed mean tells us that over 50% of the cluster peak voxels observed in a group SPM with this sample size will fail to be suprathreshold in an exact replication. Furthermore, our results represent a best case scenario for replicability (at any of our tested sample sizes for any of our tasks) because we drew samples from the same broad population. All data within each dataset were collected 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 (per dataset). 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 replicability.

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 replicability revealed small but consistent roles for several other factors, most notably the mean between-subject similarity. 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., refs. [32,33]). However, our results reinforce the measurable impact of individual differences. A long line of research has highlighted the extraordinary variability sometimes observed between individuals, and argued for taking advantage of this variability, or at least acknowledging and attempting to control for it[25–27,34–38]. Our results fit with these earlier observations that individual identity is a 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 more representative sampling.

It is possible that our results do a poor job of capturing the average replicability that should be expected across the field at large. However, we do not believe this to be the case, for four reasons. First, our results for the People Pieces Analogy (PPA) and Set-Shifting Task (SST), as well as for all Human Connectome Project (HCP) tasks, each of which separately included identical sets of participants and exactly matched pseudo-replicate groups, span a fairly wide range of replicability values (see Table 1 for a list of tasks and their definitions). Second, the results from our two distinct datasets were broadly similar, despite myriad differences at all levels of data collection and analysis. Third, the included tasks are well-known, and cover a number of cognitive domains of general interest to researchers in cognitive neuroscience. And fourth, our results are consistent with earlier work demonstrating the inadequacy of typical ($N \approx 30$) sample sizes. Although there is no simple way to map our results onto these earlier studies, the general conclusion is much the same.

Another possibility is that our results are idiosyncratic to the tasks included in our analyses. We hope that our choice of two

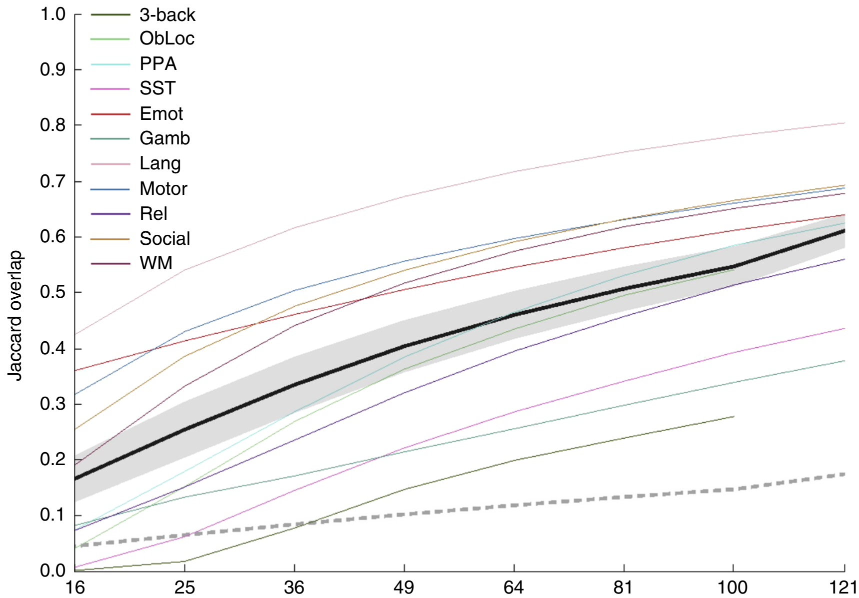

Fig. 3 Conservatively thresholded cluster-level results. Replicability results for cluster-level (thresholded conservatively) analyses, measured as the Jaccard overlap between paired group maps—that is the ratio of the number of voxels in the intersection of the two thresholded maps to the number of voxels in the union of the two. Average observed (±1 mean within-task standard deviation) shown in black (dark gray); average null (±1 standard deviation) shown in dashed medium gray (light gray). Also shown are individual task curves for each task (colors given in legend; refer to Table 1 for task abbreviations). See also Supplementary Fig. 2

datasets that span a wide range of domains yields results that are broadly informative, but in recognition of the fact that we are necessarily limited to working with a finite set of tasks, we did not attempt to draw any strong inferences from our results, e.g., regarding the sample size at which a given replicability metric was significantly non-zero. We would similarly encourage readers to avoid attempting to draw even more fine-grained inferences, for example, to assess the replicability of one task with respect to the others, or to draw inferences from the replicability of a particular task to the putative replicability of a corresponding cognitive domain; such inferences would likely require more data than we used here, and are certainly beyond the scope of the analyses we conducted.

It is also possible that results using other methods would demonstrate substantially higher replicability than the results we present. Although it is beyond the scope of the present work to adapt our approach to other analysis methods, part of our goal in using data from the HCP[39] was to enable other researchers to carry out their own analyses of these data using a similar approach, making whatever changes they see fit (in preprocessing, software tool, or analysis method).

Although our results suggest that typical sample sizes are inadequate, 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, it is clear that there is a range of replicability estimates associated with each sample size, which is due to differences in effect size. We do not focus explicitly on effect size here because, as outlined in the introduction, this concept is not easy to define in a universally meaningful way for fMRI. We do present information on the extent of activation in each task, as well as how peak height and inter-individual variability relate to replicability, each of which are presumably related to effect size. And more importantly, our tasks cover a reasonable range of effect sizes, and the impact of sample size is clear across all tasks. For readers interested in recommendations for a minimum sample size, we refer 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 accessible enough that researchers can see that typical sample sizes produce only modestly replicable results, irrespective of how replicability is measured. Thus, our results add to the growing consensus calling for a 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[34,37] or encoding methods[40], or methods which acknowledge individual variability and attempt to cluster participants into relatively homogeneous subgroups—likely deserve more attention for their potential to overcome at least one part of the problem with small samples (i.e., inter-individual variability). Of course, alternative methods are often ill-suited to answering the questions of the sort addressed by general linear model analyses; however, a paradigm shift in what kinds of questions cognitive neuroscientists focus on may be a better solution than persisting with group-based logic and having to choose between unreplicable results or studies comprising a hundred or more participants.

Replicability is the foundation of scientific progress. Unfortunately, for a variety of reasons, many scientific fields are currently gripped by an apparent crisis of irreproducibility[1]. While some of the causes of this crisis are deeply interwoven into the academic landscape—incentives related to publication, funding, and tenure—one 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[14,21,22]. Indeed, the success of quantitative meta-analysis methods (e.g.,

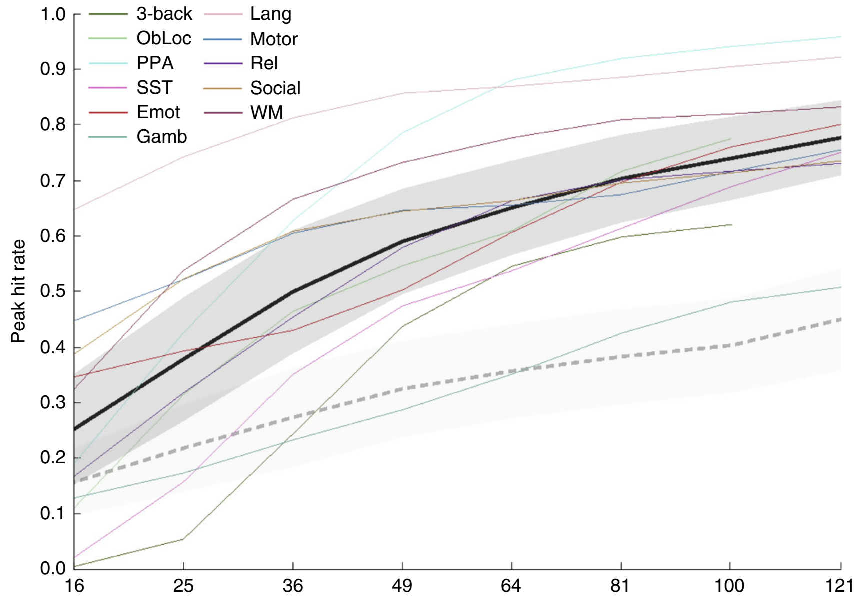

Fig. 4 Conservatively thresholded peak-level results. Replicability results for suprathreshold peak-level (thresholded conservatively) analyses, measured as the peak hit rate—that is, the proportion of cluster peaks in one map in each pair that are suprathreshold in the other map. Average observed (±1 mean within-task standard deviation) shown in black (dark gray); average null (±1 standard deviation) shown in dashed medium gray (light gray). Also shown are individual task curves for each task (colors given in legend; refer to Table 1 for task abbreviations). See also Supplementary Fig. 3

Table 1 Dataset details

Dataset	Task abbreviation	Full task name
UIUC	3-back	3-back
	ObLoc	Object Location
	PPA	People Pieces Analogy
	SST	Set-Shifting Task
HCP	Emot	Emotion Processing
	Gamb	Incentive Processing
	Lang	Language Processing
	Motor	Motor
	Rel	Relational Processing
	Social	Social Cognition (ToM)
	WM	Working Memory

Names and abbreviations of all tasks associated with each dataset

activation likelihood estimation[41]), alongside reports of moderate test–retest reliability for task-based fMRI[18], 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 replicability. Previous empirical work has demonstrated that typical fMRI sample sizes are inadequate[9,11]. Our results demonstrate that replicability (as measured at multiple levels of analysis) is quite modest at typical sample sizes, thus serving to highlight and extend these previous results.

Methods

Overview. We carried out a series of analyses across eleven distinct tasks (from two large datasets; see Table 1). Because these analyses had the same form across all eleven tasks, we describe here the details of those analyses. We first describe the details of the first dataset, followed by a description of the second dataset, then a detailed description of each task in turn, and end with a description of the analysis methods themselves.

Dataset 1 (UIUC). *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, themselves replications of refs. [42–45]). 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 2 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 227 participants were recruited for and provided data in the first intervention study (Study 1). 3-back includes a sample of 214 participants with complete data, and ObLoc includes 200 participants (of the 214 included in 3-back) with complete data.

A total of 301 participants were recruited for and provided data in the second intervention study (Study 2). For the two fMRI tasks, an identical set of 279 participants had complete data in both and are included in all analyses.

Scanning procedure: All participants in both Studies 1 and 2 were scanned on the same Siemens 3 T Magnetom Trio. Study 1 participants were scanned with a 12-channel head coil; Study 2 participants were scanned with a 32-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. ObLoc, PPA, and SST 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. 3-back 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 Task descriptions 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 (www.fmrib.ox.ac.uk/fsl) FEAT (FMRI Expert Analysis Tool, version 6.00) software package. Preprocessing included motion correction using MCFLIRT[46], BET brain extraction[47], spatial smoothing with a 6 mm full width at half maximum (FWHM) kernel, grand-mean intensity normalization, pre-whitening with the FILM tool[48], 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[46,49] and further refined using the non-linear FNIRT tool (8 mm warp resolution[50]). 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[51], and then applying the non-linear warp generated from the high-resolution scan to the functional scan.

General Linear Model analysis: For a complete description of each task, task events, and contrasts, see below. Briefly, 3-back included 7 events; ObLoc included 4 events; PPA included 7 experimental events; and SST included 10 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[52].

Dataset 2 (HCP). In addition to the data collected at University of Illinois Urbana-Champaign, we incorporated data from the HCP[39,53]. Details of the collection, preprocessing, and low- and mid-level general linear model analysis of these data can be found elsewhere (e.g., the HCP S500 Release Reference Manual: http://www.humanconnectome.org/study/hcp-young-adult/document/500-subjects-data-release/). The volumetric analysis results (smoothed with a 4 mm kernel) were downloaded from the Amazon Web Services S3 bucket designated for sharing these data (s3://hcp-openaccess) in August 2017. A total of 463 participants from this release were included in our analyses (see Supplementary Table 2 for a full list of participant IDs); the remaining participants who were included in this release but not in our analyses were excluded on the basis of QC recommendations from the HCP group or due to errors encountered during analysis. Any participant who was excluded for a single task was excluded across all tasks, such that the seven HCP tasks have an identical set of participants.

Task descriptions. 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.

3-back: See[44] 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 two, four, or five items earlier (lures). As in[44], 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).

ObLoc: See[45] 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).

PPA: See[42] 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).

SST: See[43] 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 ref. [43], 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).

HCP Tasks: See[53] for details on the tasks included in the HCP. Supplementary Table 3 lists the tasks, as well as the contrast chosen for each task (task names, contrast numbers, and contrast descriptions taken from supplemental materials of ref. [54]). We note that HCP's core research team recommends caution in the use of volumetric data; however, because our aims are orthogonal to those of most users of these task-based data, and because our measures are all designed for volumetric data, we feel that our use of these data, rather than the surface-based data, is appropriate.

Pseudo-replicate analysis. To estimate the replicability 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 replicability, 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 3-back and ObLoc, all sorts were done independently, while for PPA and SST (which comprised an identical set of participants), the same 500 sorts were applied to both tasks, and likewise, a single set of 500 sorts was applied to all HCP tasks. For the purposes of presentation, we show the average replicability estimate across all eleven tasks for each sample size, along with the average within-task (and within-sample size) standard deviation (that is, the standard deviation is computed for each task and sample size; these estimates are then averaged across tasks at each sample size), 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. 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 replicability 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, although we have access to large samples by the standards of many neuroimaging studies, they may not be 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.

Voxel-level replicability (intensity): Arguably, the goal in fMRI is to accurately capture the activity in every single voxel. Indeed, many analysis methods are predicated on the assumption that BOLD measures of voxel-level activity are meaningful, 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 replicability on which we focused was the replicability 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). The Pearson correlation gives us a holistic indication of how similar the between-voxel patterns of activity are across SPMs. 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 field of view) in either SPM.

The null distributions 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., 2nd and 98th 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 (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. We then computed the correlation between each real map and all 1000 of these null maps, and took the 95th percentile across these 1000 correlation values as the null for that specific real map; repeating this procedure across all maps yielded null values for every map, over which we took the average to arrive at the null curves presented in the figures.

Voxel-level replicability (thresholded): Without abandoning the notion of describing replicability at the voxel level, it is nonetheless possible to relax the definition of what is being replicated 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. 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, and used these thresholded maps in order to estimate the true proportion of voxels that should be suprathreshold for each task. Note that because the full samples comprise a larger number of participants than most fMRI studies—and because we will treat these full-sample results as ground truth which should be free of false positives inasmuch as possible—we have set the liberal and conservative thresholds higher than is typical in most fMRI experiments. The exact thresholds varied across data sets (in order to equate the power of the z-critical value as a function of sample size). See Table 2 for all z thresholds; all cluster p thresholds were set at $p < 0.01$.

With these per-task proportions suprathreshold (which are listed for each task in Supplementary Table 4), 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 theoretical type I or II error rates of each map, but simply the number of suprathreshold voxels. Our metric of replicability 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 (with intersection calculated subject to the constraint that the voxel have the same sign in both maps—i.e., a voxel which was positive suprathreshold in P and negative suprathreshold in Q would not count as an intersection, but this voxel would be counted in the denominator). This statistic ranges from 0 (no overlap) to 1 (perfect overlap).

The null results were generated using the same approach 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 (again, in a sign-sensitive manner) between the pair of one null and one true map. As above, this procedure was repeated 1000 times and the 95th percentile was taken for each map, and values were averaged across maps.

Cluster-level replicability: While the ultimate or idealized goal of fMRI would seem to be voxel-level replicability, 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 replicability 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 replicate a cluster. Existing methods for conducting inferential statistics on clusters (e.g., Gaussian random field theory[55]; or permutation[56],) refer to the null probability of observing a cluster of a given size (or possibly mass[57];) 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[58], from which we take inspiration for our peak-based approach, employed a liberal definition in their cluster-based methods: a cluster is declared to be 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 ($z > 1.96$, $p < 0.05$) and once at a more conservative threshold ($z > 2.81$, $p < 0.01$). As with the previous thresholding analysis, we carried these analyses out in a two-tailed fashion, running FSL's cluster-based thresholding once with $z > z_{\mathrm{crit}}$ and once with $z < -z_{\mathrm{crit}}$; then, Jaccard overlap was computed in a sign-sensitive manner.

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.

The null was computed almost identically to that described in the previous section: each null smoothed map was thresholded (two-tailed) to match the proportion of suprathreshold voxels from the corresponding true image, and the Jaccard overlap between the two was computed.

Peak-level replicability: 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[17,58]. 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 ref. [58]. Therefore, we use the same cluster-extent (Gaussian random field theory) thresholding approach as for the cluster-level analyses. That is, whereas[58] use a peak-based secondary threshold when considering peaks as topological features, we use an extent-based secondary threshold.

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 (because we are not comparing against ground truth), we can nonetheless examine the replication success of suprathreshold peaks. That is, 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 in general will 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.

As with all other thresholding-based measures, we carried this analysis out in a sign-sensitive manner—i.e., a peak from a positive cluster did not count as overlapping even if it intersected with a suprathreshold negative cluster. We used the same approach described in the preceding section to generate the null distribution. That is, we used the smoothed null maps, thresholded (two-tailed) to match proportion, to classify peaks.

Peak height replicability analysis. For our peak analyses, in contrast to our other analysis approaches, it is possible to construct a disaggregated statistic—that is, to define replicability 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, replicability, and sample size. To this end, we collated each peak z value with whether that voxel was replicated (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. Finally, we used the sim function (part of the arm package in **R**) to graphically display uncertainty around the model fits. 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(replicability) is heavily influenced by the sparsity of the counterpart map. The results of this analysis are shown in Supplementary Figs. 4, 5.

Measurables. Our expectation was that sample size would be the largest driver of replicability, 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.

Table 2 Task thresholds

Task(s)	Sample size	Liberal threshold	Conservative threshold
3-back	214	±3.50	±5.25
ObLoc	200	±3.39	±5.08
PPA, SST	279	±4.00	±6.00
All HCP tasks	463	±5.15	±7.73

Specific z thresholds used in cluster-based thresholding of full-sample analyses for use in proportion-based thresholding analysis

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.

Our modeling approach was relatively straightforward. First, we calculated per-pseudoreplicate measures for each of five variables: motion (taken as the average across functional runs of FSL's estimated mean absolute RMS per run); contrast power (average across functional runs of the reciprocal of the contrast precision $= c * \mathrm{inv}(X' * X) * c'$, where c is the contrast vector and X is the convolved design matrix); the number of outliers in the design matrix's hat matrix (average across functional runs of the count of diagonal entries on the hat matrix exceeding $2 * \mathrm{rank}(\mathrm{hat})/\mathrm{nrows}(\mathrm{hat})$, where nrows(M) is the number of rows in M; see ref. [59]); within-individual variability (the average across every pair of runs of the whole-brain correlation between the run-level SPM for the given contrast within a participant) and between-individual variability (the whole-brain correlation between individual-level SPMs for the given contrast for each pair of participants).

The first four of these measures are defined on a per-participant basis, while the last is defined at the level of participant pairs. In order to translate these measures to the pseudoreplicate level, we took the arithmetic mean of the per-participant or -participant-pair measures over all participants in a given pseudoreplicate, as well as the standard deviation. Finally, to translate these per-pseudoreplicate measures to the pseudoreplicate-pair level (which is the level at which our outcome variables are defined), we took the arithmetic mean and absolute difference between the P and Q pseudoreplicate-level measures in each pair. Thus, each of our five original variables that varied at the individual level is expanded to four variables for the purposes of modeling. The final variable, sample size, does not vary below the highest level, and so does not need to be expanded similarly. Likewise, our outcome variable of unthresholded voxelwise similarity is already defined at the appropriate level.

Once we have the explanatory variables defined for every pseudoreplicate pair for every sample size and task, we did simple task-wise ordinary least squares regression to estimate the influence of each variable. First, for each task, we removed all observations from the largest sample size per task for the outcome variable and all explanatory variables (because for some tasks for which k was near $N/2$, some of our variables had variance near zero), then demeaned all variables. Next, we orthogonalized each of the four within-subjects variability regressors with respect to the twelve preceding variables (i.e., those derived from motion, contrast power, and hat-matrix outliers), and orthogonalized each of the four between-subjects variability regressors with respect to the preceding sixteen. Finally, we regressed the outcome variable on this set of explanatory variables (removing collinear variables as needed—this only occurred for tasks for which there was no variance in contrast power across individuals, such that some of the variables derived from contrast power were undefined).

We present two measures of the strength of the relationship between each original explanatory variable and the outcome variable. The first of these measures is the effect size of the largest of the four variables derived from each original variable—that is, $\beta_i/\sqrt{(\sigma^2_{\mathrm{err}} * c_i * \mathrm{inv}(X' * X)^* c_i')}$, where σ^2_{err} is the sum of squared residuals from the model, c_i is a vector of zeros with 1 as its ith entry, and X is the design matrix. This is the standard formula for the z-value associated with a given β, without the reciprocal of the degrees of freedom in the denominator. The second measure is the increase in R^2 that results from including adding all four variables derived from one of the original explanatory variables, compared against a model that includes all variables except these four.

In order to aggregate each of these measures across tasks, we computed the maximum a posteriori estimate of each across tasks. For the measure of effect size, we used a prior distribution of (0,1) to shrink our estimated average across tasks. For the measure of ΔR^2, which ranges from 0–1, we first used a logit transform to convert the measures to a scale with infinite support, then used a prior distribution of (−20,10) in the transformed scale to shrink the estimate average toward -20 (again, in the transformed scale), and finally used the logistic transform to return the result back to the original 0–1 scale.

Note that this modeling approach is not able to probe the relationship between the explanatory and outcome variables at the level of differences between tasks. This was an intentional choice on our part because for some of the explanatory variables, between-task differences were many orders of magnitude larger than within-task differences. Moreover, because our tasks differ in many ways that aren't captured by our chosen variables, it would be extremely speculative to attribute between-task differences in replicability (with a sample of only 11 tasks) based on a model including over twenty regressors. Therefore, using an approach that relies on the consistency of the within-task relationship across tasks is conservative, although we still caution readers against drawing strong inferences from our results, because our approach is designed to be primarily descriptive.

Code availability. All custom MATLAB analysis scripts are available from https://github.com/fmrireplicability/NCB_code.

Data availability. All original data for the HCP analyses presented here are available from the AWS S3 bucket described above (s3://hcp-openaccess). The IARPA SHARP Program data repository is anticipated by Fall 2019.

Received: 7 March 2018 Accepted: 10 May 2018
Published online: 07 June 2018

References

1. Baker, M. Is there a reproducibility crisis? *Nature* **533**, 452–454 (2016).
2. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
3. Replication studies offer much more than technical details. *Nature* **541**, 259–260 (2017).
4. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
5. Szucs, D. A tutorial on hunting statistical significance by chasing N. *Front. Psychol.* **7**, 1444 (2016).
6. Barnes, R. M., Tobin, S. J., Johnston, H. M., MacKenzie, N. & Taglang, C. M. Replication rate, framing, and format affect attitudes and decisions about science claims. *Front. Psychol.* **7**, 1826 (2016).
7. Wicherts, J. M. et al. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front. Psychol.* **7**, 1832 (2016).
8. Carp, J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* **63**, 289–300 (2012).
9. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
10. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
11. Szucs, D. & Ioannidis, J. P. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, e2000797 (2017).
12. Thirion, B. et al. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage* **35**, 105–120 (2007).
13. Cremers, H. R., Wager, T. D. & Yarkoni, T. The relation between statistical power and inference in fMRI. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0184923 (2017).
14. Friston, K. Ten ironic rules for non-statistical reviewers. *Neuroimage* **61**, 1300–1310 (2012).
15. Ingre, M. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on Friston (2012). *Neuroimage* **81**, 496–498 (2013).
16. Mumford, J. A. & Nichols, T. E. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage* **39**, 261–268 (2008).
17. Durnez, J. et al. Power and sample size calculations for fMRI studies based on the prevalence of active peaks. https://www.biorxiv.org/content/early/2016/04/20/049429, 049429 (2016).
18. Bennett, C. M. & Miller, M. B. How reliable are the results from functional magnetic resonance imaging? *Ann. NY Acad. Sci.* **1191**, 133–155 (2010).
19. Gonzalez-Castillo, J. & Talavage, T. M. Reproducibility of fMRI activations associated with auditory sentence comprehension. *Neuroimage* **54**, 2138–2155 (2011).
20. Plichta, M. M. et al. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. *Neuroimage* **60**, 1746–1758 (2012).
21. Bennett, C. M. & Miller, M. B. fMRI reliability: influences of task and experimental design. *Cogn. Affect. Behav. Neurosci.* **13**, 690–702 (2013).
22. Liu, T. T., Frank, L. R., Wong, E. C. & Buxton, R. B. Detection power, estimation efficiency, and predictability in event-related fMRI. *Neuroimage* **13**, 759–773 (2001).
23. Wager, T. D. & Nichols, T. E. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *Neuroimage* **18**, 293–309 (2003).
24. Liu, T. T. & Frank, L. R. Efficiency, power, and entropy in event-related FMRI with multiple trial types: Part I: Theory. *Neuroimage* **21**, 387–400 (2004).
25. Miller, M. B. et al. Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *Neuroimage* **48**, 625–635 (2009).
26. Gabrieli, J. D., Ghosh, S. S. & Whitfield-Gabrieli, S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* **85**, 11–26 (2015).
27. Dubois, J. & Adolphs, R. Building a science of individual differences from fMRI. *Trends Cogn. Sci.* **20**, 425–443 (2016).
28. Turner, B. O. & Miller, M. B. Number of events and reliability in fMRI. *Cogn. Affect. Behav. Neurosci.* **13**, 615–626 (2013).

29. Evans, S. What has replication ever done for us? Insights from neuroimaging of speech perception. *Front. Hum. Neurosci*. https://doi.org/10.3389/fnhum.2017.00041 (2017).
30. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl Acad. Sci. USA* **113**, 7900–7905 (2016).
31. Bennett, C. M., Miller, M. B. & Wolford, G. L. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *Neuroimage* **47**, S125 (2009).
32. Rosenblatt, J. D., Vink, M. & Benjamini, Y. Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. *Neuroimage* **84**, 113–121 (2014).
33. Seghier, M. L. & Price, C. J. Visualising inter-subject variability in fMRI using threshold-weighted overlap maps. *Sci. Rep*., https://doi.org/10.1038/srep20170 (2016).
34. Van Horn, J. D., Grafton, S. T. & Miller, M. B. Individual variability in brain activity: a nuisance or an opportunity? *Brain. Imaging Behav.* **2**, 327 (2008).
35. Miller, M. B. et al. Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *J. Cogn. Neurosci.* **14**, 1200–1214 (2002).
36. Miller, M. B., Donovan, C. L., Bennett, C. M., Aminoff, E. M. & Mayer, R. E. Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *Neuroimage* **59**, 83–93 (2012).
37. Seghier, M. L. & Price, C. J. Interpreting and utilising intersubject variability in brain function. *Trends Cogn. Sci.* https://doi.org/10.1016/j.tics.2018.03.003 (2018).
38. Gratton, C. et al. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron* **98**, 439–452.e5 (2018).
39. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
40. Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. & Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902–915 (2009).
41. Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F. & Fox, P. T. Activation likelihood estimation meta-analysis revisited. *Neuroimage* **59**, 2349–2361 (2012).
42. Cho, S. et al. Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cereb. Cortex* **20**, 524–533 (2010).
43. Witt, S. T. & Stevens, M. C. fMRI task parameters influence hemodynamic activity in regions implicated in mental set switching. *Neuroimage* **65**, 139–151 (2013).
44. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6**, 316–322 (2003).
45. Hannula, D. E. & Ranganath, C. Medial temporal lobe activity predicts successful relational memory binding. *J. Neurosci.* **28**, 116–124 (2008).
46. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
47. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
48. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62**, 782–790 (2012).
49. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5**, 143–156 (2001).
50. Andersson, J. L., Jenkinson, M. & Smith, S. Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2. http://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf (2007).
51. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
52. Woolrich, M. Robust group analysis using outlier inference. *Neuroimage* **41**, 286–301 (2008).
53. Barch, D. M. et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169–189 (2013).
54. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
55. Worsley, K. J., Taylor, J. E., Tomaiuolo, F. & Lerch, J. Unified univariate and multivariate random field theory. *Neuroimage* **23**, S189–S195 (2004).
56. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
57. Zhang, H., Nichols, T. E. & Johnson, T. D. Cluster mass inference via random field theory. *Neuroimage* **44**, 51–61 (2009).
58. Durnez, J., Moerkerke, B. & Nichols, T. E. Post-hoc power estimation for topological inference in fMRI. *Neuroimage* **84**, 45–64 (2014).
59. Hoaglin, D. C. & Welsch, R. E. The hat matrix in regression and ANOVA. *Am. Stat.* **32**, 17–22 (1978).

Acknowledgements

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. We 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. We also thank Jennifer Elam and Matthew Glasser for their help in retrieving the HCP data. This work was also supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office.

Author contributions

E.J.P. and B.O.T. conducted all analyses and wrote the manuscript; M.B.M. and A.K.B. provided supervision, recommended analyses, and revised the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s42003-018-0073-z.

Competing interests: The authors declare no competing interests.