From Maps to Models: A Survey on the Reliability of Small Studies of Task-Based fMRI

Patrick Sadil,[1*] Martin A. Lindquist,[1]

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,

615 N. Wolfe Street, Baltimore, Maryland 21205, USA

*Correspondence: psadil1jh.edu

August 5, 2024

Keywords: MRI, functional MRI, task MRI, and Reproducibility

Abstract: Task-based functional magnetic resonance imaging is a powerful tool for studying brain function, but neuroimaging research produces ongoing concerns regarding small-sample studies and how to interpret them. Although it is well understood that larger samples are preferable, many situations require researchers to make judgments from small studies, including reviewing the existing literature, analyzing pilot data, or assessing subsamples. Quantitative guidance on how to make these judgments remains scarce. To address this, we leverage the Human Connectome Project's Young Adult dataset to survey various analyses–from regional activation maps to predictive models. We find that, for some classic analyses such as detecting regional activation or cluster peak location, studies with as few as 40 subjects are adequate, although this depends crucially on effect sizes. For predictive modeling, similar sizes can be adequate for detecting whether features are predictable, but at least an order of magnitude more (at least hundreds) may be required for developing consistent predictions. These results offer valuable insights for designing and interpreting fMRI studies, emphasizing the importance of considering effect size, sample size, and analysis approach when assessing the reliability of findings. We hope that this survey serves as a reference for identifying which kinds of research questions can be reliably answered with small-scale studies.

Introduction

Considerable attention has been given to problems caused by inadequate sample sizes in task-based functional magnetic resonance imaging (fMRI) studies (e.g., Button et al., 2013; M. A. Lindquist et al., 2013; Lohmann et al., 2017; Marek et al., 2022; Poldrack et al., 2017; Thirion et al., 2007; Turner et al., 2018; Yarkoni, 2009). When individual studies have too few subjects, estimated effects tend to be noisy, making the studies challenging to interpret because both false positives and false negatives rates increase (Cremers et al., 2017; Gonzalez-Castillo et al., 2012; Lieberman & Cunningham, 2009), effect sizes estimated from significant results are inflated (Marek et al., 2022; Reddan et al., 2017; Yarkoni, 2009), and the replicability of significance maps are low (Turner et al., 2018). Together these issues conspire to impede cumulative research (see also Ottenbacher, 1996; Schmidt, 1996).

To address these issues, several possible solutions have emerged, ranging from the methodological (e.g., techniques for meta-analysis, Eickhoff et al., 2009; Turkeltaub et al., 2002; Wager et al., 2003) to the social, including the organization and coordination of imaging initiatives collecting datasets comprising hundreds to tens of thousands of subjects (e.g., Di Martino et al., 2014; Krieger et al., 2017; Miller et al., 2016; Van Essen et al., 2013; Volkow et al., 2018). These solutions foreshadow good news about the reproducibility, reliability, and overall quality of research based on their use. However, it is still often necessary for researchers to interpret results from studies with relatively small sample sizes, such as when revisiting older publications and analyzing pilot data. Given the prevalence of small studies in the neuroimaging literature (Poldrack et al., 2017), it remains important to have well-calibrated expectations for their reliability.

Fortunately, data from relatively small, task-based studies can support reliable research in several specific ways (Kragel et al., 2021). For example, multivariate models trained using only tens of subjects can exhibit strong predictive performance on external samples consisting of hundreds of subjects (Han et al., 2022; Lee et al., 2021; Wager et al., 2013). If these models had been trained with larger and more diverse groups of subjects, their predictive performance would likely improve (Chen et al., 2023; Greene et al., 2022; Schulz et al., 2022; Traut et al., 2022), and small samples usually impede accurate estimation of generalization performance (Poldrack et al., 2020; Varoquaux, 2017). But many research questions can be answered based on whether, and not how well, fMRI data support predictions (Naselaris et al., 2011). When multivariate models are tested appropriately – such as through cross-validation, or, ideally, external test sets – predictive performance in a smaller dataset justifies pursuing the line of research that led to the study.

In addition to research using multivariate models, there are also examples of reliable research with relatively small samples using classic, mass-univariate approaches. As one straightforward example, high power can be achieved by focusing on robust effects (Desmond & Glover, 2002), which are prevalent in regions like the somatomotor or visual networks (Engel et al., 1994; Grodd et al., 2001). Moreover, avenues for enhancing reliability are provided by ongoing developments in statistical methods. For example, in traditional mass-univariate testing, linear models are fit to the timeseries from individual voxels, and the resulting maps can be tested for significance voxel-wise, across clusters, or regionally. With small sample sizes, results from these methods exhibit poor replicability (Nee, 2019; Turner et al., 2018, 2019). Yet several alternatives to these traditional methods exist that have been shown to improve statistical power while preserving rates of false positives and improving reliability (Noble et al., 2020; Spisák et al., 2019; Wang et al., 2021). As with multivariate methods, this success points to the fact that it may be possible to overemphasize a lack of reproducibility, and that even small sample sizes can support robust research.

In this paper, we survey the robustness of inferences that rely on task fMRI conducted on small samples, by which we mean fewer than 100 subjects. The investigation considered four approaches to analyzing task fMRI, methods that range from statistical maps to predictive models. First, we considered region-of-interest analyses, assessing the sensitivity and reliability for detecting regional activation. Second, we considered the consistency with which studies can localize peaks of activity. Third, we considered the reliability of the patterns of estimated effect sizes. Finally, we concluded with an assessment of the reliability of multivariate models and conclusions based on those models.

For each of these four approaches, we explored robustness using a four-part simulation (Cremers et al., 2017; Geuter et al., 2018; Lohmann et al., 2017; Thirion et al., 2007; Turner et al., 2018). First, we built and described a reference distribution – a population about which individual studies should support inference. Then, we simulated studies based on samples from this population across a range of sample sizes and assessed how accurately individual studies estimated population-level metrics. Third, we calculated statistics related to the variability of these metrics across simulations. Finally, we calculated the variability of these metrics across the individual subjects that compose the whole population.

Methods

Simulations relied on data from the Human Connectome Project Young Adult cohort (HCP YA; Feinberg et al., 2010; Moeller et al., 2010; Setsompop et al., 2012; Van Essen et al., 2013; Xu et al., 2012). We summarize the dataset here and describe the analyses. The tasks and contrasts included in analyses are described in the supplementary materials (section 1).

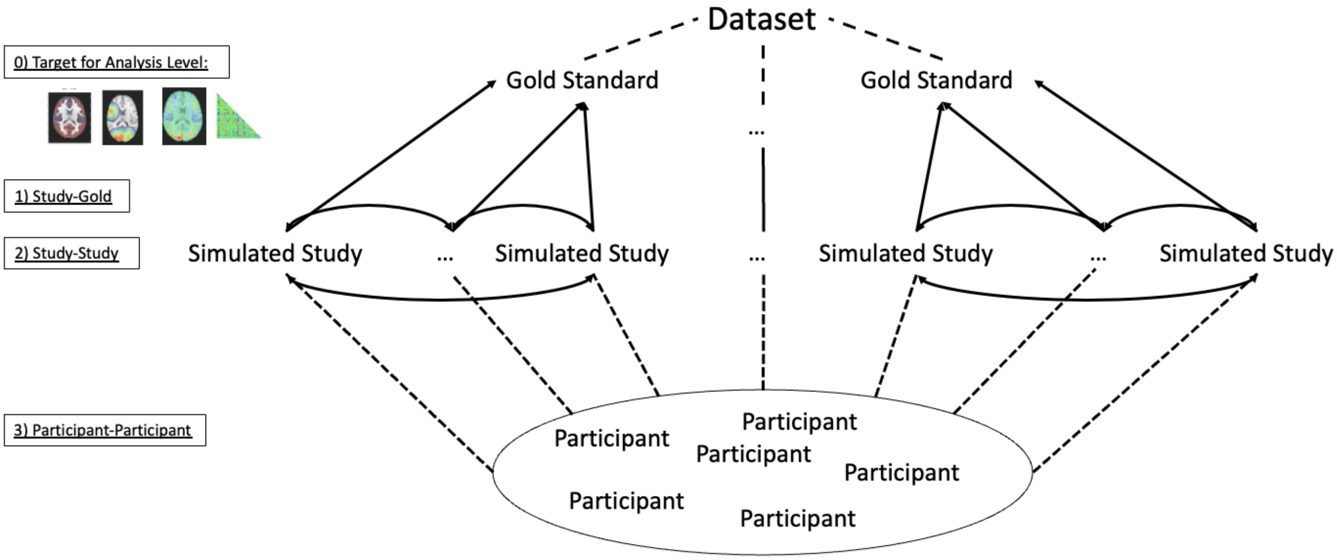

Figure 1: Outline of Methods. From a large dataset (HCP YA), gold-standard results were generated at each of four different analysis levels: regional of interest activation (Figure 2), peak localization (Figure 3), topography (Figure 4), and predictive model performance (Figure 5). Smaller studies were simulated and compared to their gold standards and each other. Finally, the participants from the gold standard were compared to each other.

Data

The Human Connectome Project 500 (HCP 500) consists of both structural and functional data from approximately 500 subjects, although the number of subjects for each task differs (numbers reported below). The 500-participant release was used because our survey covers methods performed in volumetric space.

All data were acquired on a Siemens Skyra 3T scanner at Washington University in St. Louis. For each task, two runs were acquired, one with a right-to-left and the other with a left-to-right phase encoding. Whole-brain echo-planar imaging acquisitions were acquired with a 32-channel head coil with TR=$720\,\mathrm{ms}$, TE=$33.1\,\mathrm{ms}$, flip angle=$52°$, bandwidth=$2290\,\mathrm{Hz/Px}$, in-place field-of-view=$208{\times}180\,\mathrm{mm}$, 72 slices, $2\,\mathrm{mm}$ isotropic voxels, with a multi-band acceleration factor of 8. For a full description of the fMRI data acquisition, see Van Essen et al. (2013).

Scans were preprocessed according to the HCP "fMRIVolume" pipeline (Glasser et al., 2013), which includes gradient unwarping, motion correction, fieldmap-based distortion correction, brain-boundary-based registration to a structural T1-weighted scan, non-linear registration into MNI152 space, grand-mean intensity normalization, and spatial smoothing using a Gaussian kernel with a full-width half-maximum of $4\,\mathrm{mm}$. Analyses were restricted to either a whole-brain mask or a graymatter mask that included both cortical and subcortical voxels.

Several analyses were performed using a general linear model on the volumetric contrast maps provided by the HCP (Barch et al., 2013). For each task, predictors (described for each task in the supplementary materials) were convolved with a canonical hemodynamic response function to generate regressors. To compensate for slice-timing differences and variability in the hemodynamic delay across regions, temporal derivatives were included and treated as variables of no interest. Both the data and the design matrix were temporally filtered using a linear high-pass filter (cutoff $200\,\mathrm{s}$). During model fitting, the time series was pre-whitened. For each task, a single contrast of the estimated parameters was analyzed (the result of a fixed-effects analysis on run-wise "Level 1" analyses). Note that although this approach did not include denoising strategies standard in many analysis pipelines (e.g., including motion regressors as confounds), denoising these data has been reported to not substantially improve individual-level z-statistics (Barch et al., 2013).

Analyses

For each task, one group-level dataset was calculated from all available subjects, which we alternatively refer to as the reference dataset, the gold standard, or the population dataset. Studies were simulated by drawing samples (with replacement) from the population (full) dataset. Simulated studies consisted of either 20, 40, 60, 80, or 100 subjects. At each sample size, simulations were repeated 100 times. In comparisons between participants, we considered the pairwise relationships between all participants in the population.

Maps of Regional Activation

Voxels were labeled and grouped according to either the Schaefer parcellation (after projection of the parcellation to standard space) or the Harvard-Oxford subcortical atlas (Schaefer et al., 2018). Analyses were performed using the 400-level parcellation (additional levels are presented in the supplementary materials). Regions were further grouped into one of the Yeo7 Networks or, if they were within subcortex, labeled as such (Thomas Yeo et al., 2011).

Within each region of interest, the average activity across voxels was calculated for all subjects. Activation was determined for each region by performing a t-test across subjects.

To facilitate comparisons across tasks, region of interest analyses focused on the ten regions in each task that exhibited the largest effect size in the gold standard. To compare the simulated studies to the gold standard, we calculated the proportion of studies at each sample size that exhibited significant activation in each of these ten key regions.

To compare simulated studies with each other, we calculated the mean square contingency coefficients, Φ, across all pairs of significance maps at a given sample size and task (thresholded at $p < 0.05$, family-wise error-corrected across all regions in the parcellation; Holm, 1979). This measure was chosen to equally prioritize each of the quadrants of the confusion matrix for statistical significance. Referring to the counts in those quadrants as True Negative, TN, True Positive, TP, False Negative, FN, and False Positive, FP

$$\Phi = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Finally, to compare subjects with each other, we calculated the product-moment correlation across subjects' regional activation maps, which were defined as the average (across voxels within a region) contrast of the parameter estimate.

Localization of Peaks Across Voxels

Reference peaks were based on the raw z-statistic map obtained using the full dataset (i.e., all available HCP subjects). The peaks were defined as those voxels that had a more extreme value than their 26 neighbors, calculated using `cluster` from FSL (S. M. Smith et al., 2004).

Peaks in simulated studies were identified in unsmoothed, probabilistic threshold-free cluster enhanced z-maps (Spisák et al., 2019). Before peak detection, maps were thresholded according to a statistical significance filter (family-wise error rate $p < 0.05$; Spisák et al., 2019). As this thresholding removes voxels with negative activation, displays that include peak location only include positive effects. Note not every simulated study included suprathreshold voxels (gambling with 20 subjects: 4 simulations with no voxels, relational with 20 subjects: 43, social with 20 subjects: 1, relational with 40 subjects: 4).

Analogous to the region of interest analyses, we considered only a subset of the peaks, to facilitate comparisons across tasks. Specifically, we considered the ten peaks that had the largest (positive) activation in the gold standard, taking at most one peak per parcel or region (the largest). These were compared to the simulated studies by calculating the proportion of stimulated studies that contained any local peak within different radii of a sphere centered on the reference peak (radii: 2, 4, 8, 10, 20 mm).

Comparisons between studies and between subjects involved associating the peaks from the z-maps of individual studies or the z-maps from individual subjects that were closest to the ten reference peaks. Comparisons were made by assessing distributions of peak Euclidean distances.

Reliability of Maps of Effect Sizes

Voxel-wise effect sizes were measured using Cohen's d. To construct the reference map, for voxel v, effect sizes were calculated from the average (across subjects), μ_v, and standard deviation (also across subjects), σ_v, of each contrast of parameter estimate (i.e., $d_v = {}^{\mu_v}/_{\sigma_v}$). Voxels were binned into categories using the guidelines of Cohen (1988): with "negligible" indicating $|d_v| < 0.2$, "small" indicating $0.2 <= |d_v| < 0.5$, "medium" indicating $0.5 <= |d_v| < 0.8$, and "large" indicating $0.8 <= |d_v|$. The effect sizes in simulated studies were calculated using Hedges' correction (Bossier et al., 2019; Hedges, 1981).

To compare the gold standard with the simulated studies, rank correlations were calculated between the maps of the reference and the simulated studies, within a gray matter mask. To compare simulated studies with each other, analogous rank correlations were calculated pairwise, between simulated studies (using effect sizes), and likewise for comparisons between subjects (using their contrasts of parameter estimates).

Reliability of Models Trained with Maps of Effect Sizes

Models predicting participant traits were trained using a form of connectome-based predictive modeling (Gao et al., 2019; Greene et al., 2018). First, the data were subjected to basic cleaning: linear detrending, band-pass filtering (0.01–$0.1\,\mathrm{Hz}$), voxel-wise standardization, and nuisance regression with 24 motion parameters (Friston et al., 1996; Satterthwaite et al., 2013). To build connectivity matrices, the left-right and right-left timeseries for each task were concatenated and then decomposed according to the 64 region version of the DiFuMo atlas (Dadi et al., 2020). Connectivity was estimated via a regularized procedure (Ledoit & Wolf, 2004), and the correlations were transformed with the inverse hyperbolic tangent. As our outcome we used several instruments (e.g., task performance and fluid intelligence); see Figure S8 for a complete list.

Considering the small sample sizes and high numbers of features (around 500 subjects but 2079 features), models relied on feature selection and regularization. Feature engineering and modeling were performed with `scikit-learn` (Pedregosa et al., 2011). First, features with a variance of less than 0.01 were removed. Then, features were independently standardized using a robust normalization (i.e., removing the median and scaling by the interquartile range). For predictions, A ridge regression model was used where the regularization parameter was selected from 20 log-spaced values (0.1 to 10000, inclusive) by the efficient leave-one-out cross-validation procedure described by Rifkin and Lippert (2007) and implemented in scikit-learn.

To facilitate comparison across studies of various sizes (including with the gold standard), a test set was segregated from the training and validation samples. The test set comprised $20\,\%$ of the full dataset (the same test set was used for all simulated studies of a given task).

Model performance was measured as a rank correlation between trained model predictions and the true values in the test set. The gold standard was defined as the performance of the model on the test sample when the training sample comprised all other subjects. This gold standard was compared to the performance of the simulated studies of varied sample sizes. At each sample size, performance was measured via a random-effects meta-analysis of the correlations estimated at each sample size as implemented in the R package meta (R Core Team, 2023; Schwarzer et al., 2015). To compare performance across studies, the predictions on the held-out test set were used to calculate an intraclass correlation. We calculated both consistency and agreement based on a 2-way random-effects model. Given that we were interested in the reliability of a single study, we used the single-measure version. These measures are also known as ICC(A,1) or ICC(C,1) (McGraw & Wong, 1996). Intraclass correlations were calculated using the R package irr (Gamer et al., 2019; R Core Team, 2023). Subjects were compared to each other by calculating the pairwise rank correlation of each connectivity feature.

Results

Maps of Regional Activation

Gold Standard

We first explored the statistical power for detecting regional activation at different sample sizes. For each task, we designated a set of regions as the “primary targets” if their average effect size in the gold standard was among the ten highest. This designation enabled us to focus on a core set of regions for each task. These regions can be understood as representing regions that would likely be studied with the given task (for region definition, see Methods). For example, in the motor task, this scheme picks voxels within the primary motor cortex.

Recovery of Gold Standard

Studies using the language and motor tasks would be nearly guaranteed to detect an effect in all ten of the targeted regions even with only 20 subjects (Figure 2a). In the social and working memory tasks, similarly high power would require around 40 subjects. Finally, the gambling and relational task would require between 60 to 80 subjects. Note that the high power is attained while maintaining family-wise error rate correction (Figure S1). These patterns were largely consistent across different levels of parcellation granularity (Figure S2).

Study-to-Study Comparisons

To compare studies, we calculated the distribution of mean square contingency coefficients across all pairs of studies at given sample sizes (Figure 2b). Tasks with the most reliably activated primary target regions had the highest distributions of coefficients. In particular, the language task exhibited a median mean square contingency coefficient with 40 subjects that was higher than the equivalent median of the gambling studies conducted with 100 subjects (0.69 vs 0.51), indicating substantially more variability in which regions were activated during any given gambling study.

Subject-to-Subject Comparisons

Finally, to compare subjects we calculated the pairwise product-moment correlation between average regional effect sizes (Figure 2c). The two most extreme tasks were again gambling and language, with the average correlation in the gambling task being 0.05 and the average correlation in the language task around 0.51. This indicates substantially more variability across subjects in which regions were activated while performing the gambling task. This may help explain the subsequent variability observed across studies (simulations).

Localization of Peaks Across Voxels

Gold Standard

Second, we considered the replicability of peak location. Due to the number of subjects, clusters of activation covered most of the gray matter, so we focused on local rather than global cluster peaks. As before, we assume that each task was designed to elicit activation in several distinct regions, but that the number of regions may vary across tasks. To facilitate comparisons across tasks, we considered the ten voxels with the highest peaks for each contrast. Note that this classification does not guarantee that the peak voxels are in one of the regions with the largest average effect size.

Recovery of Gold Standard

To summarize the replicability in a manner that was independent of the atlas, we considered the distance between the highest reference peaks and the nearest peak within the simulated studies. Specifically, we calculated the proportion of simulated studies that contained a significant peak within various radii, plotting these proportions by task and sample size (Figure 3a). Increasing the number of subjects improved peak localizability (e.g., at a given radius for a given task, the proportions increase down the rows of Figure 3a). In all tasks except gambling and relational, nearly $100\,\%$ of simulated studies were within $10\,\mathrm{mm}$ of each of the ten peaks from the gold standard – even with only 20 subjects. For the gambling and relational tasks, simulated studies with 20 subjects often failed to provide a peak within that radius. Even so, with 40-60 subjects in these two tasks, over $95\,\%$ of simulated studies produced peaks that were within $10\,\mathrm{mm}$ (equivalently, five voxels) of one of the top ten peaks.

When considering all peaks within each contrast (that is, all peaks with effect sizes above negligible magnitude and in voxels that survived multiple familywise error corrections), localization exhibited a dependence on the effect size of the peak such that peaks with larger effects in the reference distribution were better localized with fewer subjects (Figure S4). For example, in studies of 20 subjects, peaks in the reference sample with small effect sizes were separated from the study peaks by an average of $34.6\,\mathrm{mm}$, while the same average for peaks whose voxels had a large effect was only $4.45\,\mathrm{mm}$.

There were apparent differences in localizability when categorizing peaks according to connectivity network (Thomas Yeo et al., 2011), such that peaks within "lower-level" networks like the somatosensory or visual networks were localized more easily than those within "higher-order" networks like the default or limbic networks (Figure S3). However, there was also a close relationship between the presence of a peak within a network and the height of the peak, so the effect of the network was not necessarily distinct from the effect of peak height. As one example, consider that, with only 20 subjects, peaks within the somatomotor network were localized well (mean: $4.6\,\mathrm{mm}$) in the motor task (average effect size: 0.84), but peaks within that network were localized relatively poorly (mean: $24.8\,\mathrm{mm}$) in the social task (average effect size: 0.48).

Study-to-Study Comparisons

To compare studies, the local peaks associated with the ten highest peaks were grouped, and the average distance between them was calculated (Figure 3b). This measure highlights the variability that can be expected across studies conducted with small samples. For example, $10\,\%$ of simulated studies with 20 subjects that used the gambling and relational tasks had peaks that were over $15\,\mathrm{mm}$ apart. In contrast, $90\,\%$ of 20 participant studies using the motor task were separated by less than $3.75\,\mathrm{mm}$.

Subject-to-Subject Comparisons

To compare subjects with each other, the study-to-study comparison was repeated with the participant-wise z-maps for each contrast, thresholding only to exclude values below 0 (Figure 3b). This revealed that the peaks in the participant maps that were closest to the gold-standard maps tended to be within a few voxels of each other. In addition, this distance was largely consistent across tasks.

Reliability of Maps of Effect Sizes

Gold Standard

Peaks are only one feature of the activation map, and for these next analyses, we reviewed the reproducibility of the map as a whole, using voxel-wise effect sizes.

For most tasks, the standard deviation of the voxel-wise variability increases with the mean, resulting in a funnel-shaped pattern (Figure 4a). The contrasts for the gambling and relational tasks had distributions with the smallest averages, resulting in the largest proportion of voxels with negligible effects and the smallest proportion of voxels with medium and large effects (Figure S5). For the gambling task in particular, fewer than $1\,\%$ of voxels had an effect size that was medium or large. In the language, motor, social, and working memory tasks, small effects were present in around 30 to $40\,\%$ of voxels, and medium effects were present in 10 to $25\,\%$. In most tasks, large effects were present in fewer than $5\,\%$ of voxels. But in the language task, a large effect was present in almost $20\,\%$ of voxels.

Recovery of Gold Standard

As measured by correlations, individual studies were able to recover the gold standard well. For all tasks, $99\,\%$ of rank correlations between the effect size maps of the simulated studies and the reference map were above 0.5 (Figure 4b). Consistent with the gambling task eliciting smaller effects, the correlations for this task were generally lower. In contrast, the language task, which tended to elicit the strongest activation, showed correlations that were typically above 0.75 at all sample sizes. These trends were also present in the correlations between simulated studies (Figure 4c); for example, all of the maps from simulated studies with only 20 subjects with the language contrast exhibited correlations above 0.8, whereas none of the maps from simulated studies with the gambling contrast reached that level, even in studies of 100 subjects.

As with the reliability of activation and peak localization, there was variation in the recovery of the gold standard across networks, and that variation was consistent with an important role for the effect sizes within the networks (Figure S6). Across all tasks, voxels within the limbic network were among those that exhibited the lowest correlations. In most tasks, voxels within the subcortical regions also exhibited low correlations. Correlations for voxels within the somatomotor network were neither the highest nor lowest for all tasks except the motor task, where they were the highest.

Study-to-Study Comparisons

The trends for the recovery of the gold standard were present in comparisons between the simulated studies, though the magnitudes were generally lower (Figure 4c). The gambling task exhibited the lowest correlations, ranging from around 0.2 with 20 subjects to .55 with 100. In contrast, the language task exhibited the highest correlations, ranging from around 0.9 with 20 subjects to .97 with 100 subjects.

Subject-to-Subject Comparisons

Correlations between the effect size maps of individual subjects were lower than those between studies, although the ordering across tasks was consistent (Figure 4d). In the gambling task, the average correlation was slightly above zero. In contrast, the average participant-to-participant correlation in the language task was nearly 0.4 but with a long left tail. The spread in the working memory task was the largest, with the 10% and 90% quantiles ranging from 0.01 to 0.30 (compare with relational, which ranged from -0.03 to 0.18).

Reliability of Models Trained with Maps of Effect Sizes

Gold Standard

Models were trained to predict characteristics of subjects within the HCP. In the main text, We focus on prediction of fluid intelligence, but results for other measures are presented in the Supplementary Materials (Figure S7). Predictions were based on features derived from connectivity matrices and a ridge regression model.

Predictions for fluid intelligence were significant with all of the considered tasks and contrasts ($p < 0.05$, as measured with permutation test for rank correlations on held-out test data, Figure 5a). The rank ordering of all tasks was similar to the other analyses, with the relational and gambling tasks exhibiting the lowest rank correlations (0.23 and 0.27), and the language and social tasks exhibiting the highest (0.35 and 0.47).

Recovery of Gold Standard

Study reliability was assessed by comparing levels of predictive performance attained with the simulated study to the performance attained with the population. All tasks were numerically below gold-standard level performance with even 100 subjects (Figure 5a). Although absolute performance was below the gold standard, the $95\,\%$ confidence intervals in all tasks excluded 0, even with only 20 participants, indicating that the ability to predict fluid intelligence can be reliably detected with even a small sample size.

Study-to-Study Comparisons

To measure the reliability of predictions across studies, the intraclass correlation coefficient was used. This enables asking about the stability of predictions for a given subject across training datasets – by how much do predictions vary according to the training set. Across all tested sample sizes and tasks, the reliability was generally poor, with intraclass correlations below 0.5 (Figure 5a). Note that this poor reliability was not limited to this particular instrument; of the instruments considered, the chosen measure, fluid intelligence, exhibited among the highest levels of reliability (Figure S8 and Figure S9).

Subject-to-Subject Comparisons

Subjects were compared to each other by calculating correlations of features in a pairwise manner. Across tasks, there was minimal difference in the distribution of pairwise correlations (Figure 5c). The relational task had the lowest $5\,\%$ quantile at 0.24 ($95\,\%$: 0.56), and the working memory task had the highest at 0.32 ($95\,\%$: 0.62).

Discussion

In this report, we leveraged the Human Connectome Project to survey several aspects of sensitivity and reliability in group-level MRI studies conducted with sample sizes common in the neuroimaging literature – typically less than 100. We used the rich datasets provided by the HCP to construct reference datasets and asked how well features of those datasets could be recovered from individual studies. We considered how these analyses were influenced by sample size, effect size, and task, aiming to provide results that could be used to calibrate expectations about the reliability of individual studies and analyses.

First, we considered the case of a researcher using the classical, univariate approach of detecting activation in task-related regions. Predictably, regions with large effect sizes could be detected with relatively small sample sizes – around 40 subjects (Figure 2a). For reference, note that reaching $80\,\%$ power with a one-sample, two-sided t-test with a true effect size of 0.8 requires at least 15 observations, which increases to 34 observations with an effect size of 0.5, and with an effect size of 0.2, 199 observations are required. However, large effect sizes were relatively rare and nearly absent from some tasks (Figure 4a, Figure S5). This means that reports of novel effect effects – even large ones – that are based on only 40 subjects should be viewed with skepticism (see also Marek et al., 2019; Reddan et al., 2017). That is, while observation of a significant effect with at least 40 subjects provides justification for deciding whether to continue a line of research, the relative scarcity of large effects underscores a need for replication before effects are considered established.

When considering peak activation, we observed that studies with between 40 - 60 subjects were able to localize a peak to within $10\,\mathrm{mm}$. For reference, note that the number of distinct regions within the cortex has been estimated to be around 300 - 400 (Van Essen et al., 2012). At this scale, the average volume of a region is approximately $2400\,\mathrm{mm}^3$. A spherical region with this volume would have a radius of around $8.3\,\mathrm{mm}$. This implies that, with 40 subjects, individual studies would contain local peaks that are within one to two "regions" away from the peaks in the gold standard. That resolution may be inadequate for certain kinds of experimental questions, especially those based on the parcellation or segmentation of microstructures (e.g., identifying substructures within subcortex). Moreover, given that analyses were performed in volumetric space, even a distance of a few millimeters could place peaks in functionally distinct cortex (e.g., spanning multiple gyri). However, considering that the voxel size in this dataset was $2.4\,\mathrm{mm}^3$, a distance of $10\,\mathrm{mm}$ translates to separation by only a few voxels.

Note that there is no guarantee that the peak activation for all voxels will be in the same location for each participant; due to idiosyncrasies across subjects (or inadequate spatial normalization), some variability in peak location may be unavoidable. Put another way, the location of the peak in the population should not be taken to be the location of a peak in any individual participant. For the tasks and contrasts studied here, most subjects exhibited peaks in activation that were less than 6mm from each other (Figure 3c). The success of methods that aggregate information across subjects using functional information (e.g., hyperalignment; Haxby et al., 2011), implies that performing diffeomorphic transformations based on structural features alone can result in misaligned features. However, based on the variability in peaks across the region of activation, the population provides a good summary of individual subjects.

Peaks are only one summary of activation. It has previously been reported that most patterns of activation produced by the Human Connectome Project are diffuse (Cremers et al., 2017), and when the activation is diffuse – especially when a cluster of active voxels span multiple anatomically defined regions – then the relevance of a peak is less clear, and localizing a peak to within $10\,\mathrm{mm}$ may be sufficient. When considering the topography of the statistical maps as a whole, correlations between individual studies and their respective gold standards ranged from strong to very strong (0.5 - 1). For tasks that elicit strong activation, maps constructed with only 20 subjects correlated with the gold standard at over 0.9, a map constructed from over 400 subjects. That is, with respect to this global measure, 20 subjects provided information that appears to be highly predictive of the maps produced by 400 subjects.

Compare these whole map correlations to those reported by the Neuroimaging Analysis and Replication Project (Botvinik-Nezer, 2020). In that project, teams of researchers analyzed the same set of data, each using the idiosyncratic set of methods preferred by individual teams. One conclusion from that replication project is that the analysis pipeline has a strong influence on binary activation maps (see also Bowring, Maumet, & Nichols, 2019), but a substantially weaker influence on the underlying statistical maps. That is, when multiple analysis pipelines are applied to a common dataset, the unthresholded statistical maps are largely consistent with each other (M. Lindquist, 2020; Taylor et al., 2023). Similarly, we show that when the same analysis pipeline is applied to several repeated experiments, the unthresholded statistical maps are again largely consistent with each other.

When interpreting the results of voxel-wise effect sizes, it may be helpful to consider “worst-case” scenarios. For small effects, the 0.05 and 0.95 quantiles for the difference between the estimated and true effects were -0.37 and 0.40 with 20 subjects, -0.24 and 0.26 for 50 subjects, and -0.18 and 0.19 with 100 subjects. That is, with 20 subjects, it is likely that the difference between the reported and true effect sizes may be off by as much as a "small" effect, whereas with 100 subjects that difference is often "negligible". Although 100 subjects exceed what can be collected for most individual studies, this scale is typical for datasets provided by many collaborative efforts. The stability at 100 subjects supports the use of datasets of this size in, for example, pilot analyses for designing more focused studies using a smaller number of subjects.

Finally, the results using multivariate models were mixed. In general, tens of subjects were sufficient for obtaining significant predictions on external training samples (Figure 5a). However, the consistency and agreement of these predictions were poor (Figure 5a, Figure S8, Figure S9). This means that the parameters learned by models remained unstable at these sample sizes. We speculate that this instability may relate to the high imbalance between the number of subjects (in the tens) and the number of features (in the thousands). The models used regularization (ridge regression), but the particular regularization procedure aims to aid cross-validation performance and not necessarily feature consistency. Data may comprise multiple, possibly disjoint, sets of features that can support equally good predictions, and so without additional information, the regularization procedure can be expected to select different sets of features across studies.

Recommendations

First, we continue to remind neuroimaging researchers that, when feasible, data ought to be made publically available. There have been calls for open sharing for over a decade. Although tools have been developed to work around the lack of easily available raw images or statistical maps (e.g., neurosynth.org) and community-driven efforts demonstrate the feasibility of sharing original data (e.g., the FCON 1000 project), there now exist many resources that obviate these workarounds. Based in the US alone, resources include OpenNeuro, the National Institute of Mental Health Data Archive, and NeuroVault. Together, these make it feasible for many more researchers to share data (Gorgolewski et al., 2015). Having rich and varied raw data drastically increases the value of small studies, especially when analyses are exploratory or focused on subtle effects.

Second, for certain well-circumscribed aims, we recommend against overemphasizing lack of reproducibility; datasets with tens of subjects, which are typical in the neuroimaging literature (Poldrack et al., 2017), can be of high quality and value. That is, for well-studied tasks that produce large effect sizes like the language, motor, social, and working memory tasks of the HCP dataset, 40 subjects provide high power to detect regional activation. But we emphasize that this recommendation applies to datasets that are as high-quality as the HCP, with tasks that are known to produce at least medium effect sizes, and with limited room for exploratory analyses. In tasks with unverified effect sizes, tens of subjects may be too few to have confidence in a new effect, or in which regions are most activated by the tasks. Even so, the sample sizes required for drawing reliable conclusions from these tasks may not be in the hundreds, considering that around 80 subjects were enough to reliably activate the targeted regions with even the gambling and relational tasks.

Third, there is a need for further research on quantifying confidence in peak location. The distributions of peak distances that we report provide heuristics for assigning confidence to locations reported in individual studies, but these heuristics imply a general lack of confidence (e.g., with 40 subjects, any voxel within $10\,\mathrm{mm}$ of a reported peak is a likely location for the true peak). Advances have been made in exploring confidence in effect size maps (Bowring, Telschow, et al., 2019; Bowring et al., 2021), but these methods are not yet commonly used. Typical cluster analyses discard substantial information about the location of activation; the significance of a cluster only implies that there is an activation in some voxel within a cluster (Woo et al., 2014). This can lead to situations where larger study populations increase the power to detect activation within each voxel, thereby increasing the size of clusters and hindering the determination of which voxels are active (Rosenblatt et al., 2018). To some extent, our reliance on pTFCE specifically can be expected to mitigate the problem of "cluster leakage" (Spisák et al., 2019). However, an issue remains where testing for activation and then selecting the voxel with the largest activation does not necessarily provide information about whether that voxel is significantly higher than neighboring voxels; a voxelwise test is still about whether each voxel's intensity is different than 0 rather than about differences between voxels.

Finally, we caution against using predictive models that have been trained with tens of subjects in any applied setting. This recommendation derives from the study-to-study comparisons of model predictions (Figure 5b), which revealed poor consistency and agreement across training datasets. That is, although models may achieve significant, even substantial, performance when trained on only tens of subjects, the specific predictions are not stable across training sets with under 100 subjects.

Data and Code Availability

Code to reproduce analyses is available on GitHub: https://github.com/psadil/maps-2-models. Analyses relied on open data provided by the Human Connectome Project, which can be downloaded from the HCP website https://humanconnectome.org/study/hcp-young-adult/document/500-subjects-data-release. Analysis results are available at https://doi.org/10.5281/zenodo.12686151.

Author Contributions

Patrick Sadil: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Martin A. Lindquist**: Conceptualization, Methodology, Validation, Formal Analysis, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

Funding

This work was supported by R01 EB026549 from the National Institute of Biomedical Imaging and Bioengineering and R01 MH129397 from the National Institute of Mental Health.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Statement

Informed consent was obtained from all Human Connectome Project Subjects.

Acknowledgements

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., & Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., Grabowski, T. J., Langfitt, J. T., Loring, D. W., Lowe, M. J., Koenig, K., Morgan, P. S., Ojemann, J. G., Rorden, C., Szaflarski, J. P., Tivarus, M. E., & Weaver, K. E. (2011). Mapping anterior temporal lobe language areas with fMRI: A multicenter normative study. *NeuroImage*, *54*(2), 1465–1475. https://doi.org/10.1016/j.neuroimage.2010.09.048

Bossier, H., Nichols, T. E., & Moerkerke, B. (2019, December). *Standardized Effect Sizes and Image-Based Meta-Analytical Approaches for fMRI Data* (Preprint). Neuroscience. https://doi.org/10.1101/865881

Botvinik-Nezer, R. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*, 4.

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Human Brain Mapping*, *40*(11), 3362–3384. https://doi.org/10.1002/hbm.24603

Bowring, A., Telschow, F., Schwartzman, A., & Nichols, T. E. (2019). Spatial confidence sets for raw effect size images. *NeuroImage*, *203*, 116187. https://doi.org/10.1016/j.neuroimage.2019.116187

Bowring, A., Telschow, F. J., Schwartzman, A., & Nichols, T. E. (2021). Confidence Sets for Cohen's d effect size images. *NeuroImage*, *226*, 117477. https://doi.org/10.1016/j.neuroimage.2020.117477

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., & Yeo, B. T. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(5), 2322–2345. https://doi.org/10.1152/jn.00339.2011

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Castelli, F., Happé, F., Frith, U., & Frith, C. (2013). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. In *Social neuroscience* (pp. 155–169). Psychology Press.

Chen, Z., Hu, B., Liu, X., Becker, B., Eickhoff, S. B., Miao, K., Gu, X., Tang, Y., Dai, X., Li, C., Leonov, A., Xiao, Z., Feng, Z., Chen, J., & Chuan-Peng, H. (2023). Sampling inequalities affect generalization of neuroimaging-based diagnostic classifiers in psychiatry. *BMC Medicine*, *21*(1), 241. https://doi.org/10.1186/s12916-023-02941-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI (S. Gilbert, Ed.). *PLOS ONE*, *12*(11), e0184923. https://doi.org/10.1371/journal.pone.0184923

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., & Mensch, A. (2020). Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, *221*, 117126. https://doi.org/10.1016/j.neuroimage.2020.117126

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the Hemodynamic Responses to Reward and Punishment in the Striatum. *Journal of Neurophysiology*, *84*(6), 3072–3077. https://doi.org/10.1152/jn.2000.84.6.3072

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, *118*(2), 115–128. https://doi.org/10.1016/S0165-0270(02)00121-8

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., . . . Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Drobyshevsky, A., Baumann, S. B., & Schneider, W. (2006). A rapid fMRI task battery for mapping of visual, motor, cognitive, and emotional function. *NeuroImage*, *31*(2), 732–744. https://doi.org/10.1016/j.neuroimage.2005.12.016

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, *30*(9), 2907–2926. https://doi.org/10.1002/hbm.20718

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., & Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature*, *369*(6481), 525–525. https://doi.org/10.1038/369525a0

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., Miller, K. L., Ugurbil, K., & Yacoub, E. (2010). Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging (P. A. Valdes-Sosa, Ed.). *PLoS ONE*, *5*(12), e15710. https://doi.org/10.1371/journal.pone.0015710

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fmri time-series. *Magnetic resonance in medicine*, *35*(3), 346–355.

Gamer, M., Lemon, J., & ¡puspendra.pusp22@gmail.com¿, I. F. P. S. (2019). *Irr: Various coefficients of interrater reliability and agreement* [R package version 0.84.1]. https://www.r-project.org

Gao, S., Greene, A. S., Constable, R. T., & Scheinost, D. (2019). Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage*, *201*, 116038. https://doi.org/10.1016/j.neuroimage.2019.116038

Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. A. (2018, April). *Effect Size and Power in fMRI Group Analysis* (Preprint). Neuroscience. https://doi.org/10.1101/295048

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gonzalez-Castillo, J., Saad, Z. S., Handwerker, D. A., Inati, S. J., Brenowitz, N., & Bandettini, P. A. (2012). Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proceedings of the National Academy of Sciences*, *109*(14), 5487–5492. https://doi.org/10.1073/pnas.1121049109

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., Yarkoni, T., & Margulies, D. S. (2015). NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00008

Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, *9*(1), 2807. https://doi.org/10.1038/s41467-018-04920-3

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., Tokoglu, F., Spann, M. N., Carrión, C. I., Barron, D. S., Sanacora, G., Srihari, V. H., Woods, S. W., Scheinost, D., & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, *609*(7925), 109–118. https://doi.org/10.1038/s41586-022-05118-w

Grodd, W., Hülsmann, E., Lotze, M., Wildgruber, D., & Erb, M. (2001). Sensorimotor mapping of the human cerebellum: fMRI evidence of somatotopic organization: Sensorimotor Mapping of the Cerebellum. *Human Brain Mapping*, *13*(2), 55–73. https://doi.org/10.1002/hbm.1025

Han, X., Ashar, Y. K., Kragel, P., Petre, B., Schelkun, V., Atlas, L. Y., Chang, L. J., Jepma, M., Koban, L., Losin, E. A. R., Roy, M., Woo, C.-W., & Wager, T. D. (2022). Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *NeuroImage*, *247*, 118844. https://doi.org/10.1016/j.neuroimage.2021.118844

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, *72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

Krieger, D., Shepard, P., Zusman, B., Jana, A., & Okonkwo, D. O. (2017, December). Shared High Value Research Resources: The CamCAN Human Lifespan Neuroimaging Dataset Processed on the Open Science Grid.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Lee, J.-J., Kim, H. J., Čeko, M., Park, B.-y., Lee, S. A., Park, H., Roy, M., Kim, S.-G., Wager, T. D., & Woo, C.-W. (2021). A neuroimaging biomarker for sustained experimental and clinical pain. *Nature Medicine*, *27*(1), 174–182. https://doi.org/10.1038/s41591-020-1142-7

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*(4), 423–428. https://doi.org/10.1093/scan/nsp052

Lindquist, M. (2020). Neuroimaging results altered by varying analysis pipelines. *Nature*, *582*(7810), 36–37. https://doi.org/10.1038/d41586-020-01282-z

Lindquist, M. A., Caffo, B., & Crainiceanu, C. (2013). Ironing out the statistical wrinkles in "ten ironic rules". *NeuroImage*, *81*, 499–502. https://doi.org/10.1016/j.neuroimage.2013.02.056

Lohmann, G., Stelzer, J., Müller, K., Lacosse, E., Buschmann, T., Kumar, V., Grodd, W., & Scheffler, K. (2017, March). *Inflated false negative rates undermine reproducibility in task-based fMRI* (Preprint). Neuroscience. https://doi.org/10.1101/122788

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., . . . Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Marek, S., Tervo-Clemmens, B., Nielsen, A. N., Wheelock, M. D., Miller, R. L., Laumann, T. O., Earl, E., Foran, W. W., Cordova, M., Doyle, O., Perrone, A., Miranda-Dominguez, O., Feczko, E., Sturgeon, D., Graham, A., Hermosillo, R., Snider, K., Galassi, A., Nagel, B. J., . . . Dosenbach, N. U. (2019). Identifying reproducible individual differences in childhood functional brain networks: An ABCD study. *Developmental Cognitive Neuroscience*, *40*, 100706. https://doi.org/10.1016/j.dcn.2019.100706

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., . . . Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*(5), 1144–1153. https://doi.org/10.1002/mrm.22361

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 130. https://doi.org/10.1038/s42003-019-0378-6

Noble, S., Scheinost, D., & Constable, R. T. (2020). Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. *NeuroImage*, *209*, 116468. https://doi.org/10.1016/j.neuroimage.2019.116468

Ottenbacher, K. J. (1996). The Power of Replications and Replications of Power. *The American Statistician*, *50*(3), 271. https://doi.org/10.2307/2684673

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, *77*(5), 534. https://doi.org/10.1001/jamapsychiatry.2019.3671

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reddan, M. C., Lindquist, M. A., & Wager, T. D. (2017). Effect Size Estimation in Neuroimaging. *JAMA Psychiatry*, *74*(3), 207. https://doi.org/10.1001/jamapsychiatry.2016.3356

Rifkin, R. M., & Lippert, R. A. (2007). *Notes on Regularized Least Squares* (tech. rep. No. MIT-CSAIL-TR-2007-025).

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., & Goeman, J. J. (2018). All-Resolutions Inference for brain imaging. *NeuroImage*, *181*, 786–796. https://doi.org/10.1016/j.neuroimage.2018.07.060

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, *64*, 240–256.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. https://doi.org/10.1037/1082-989X.1.2.115

Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D., & Ritter, K. (2022, February). *Performance reserves in brain-imaging-based phenotype prediction* (Preprint). Neuroscience. https://doi.org/10.1101/2022.02.23.481601

Schwarzer, G., Carpenter, J. R., Rücker, G., et al. (2015). *Meta-analysis with r* (Vol. 4784). Springer.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced *g* -factor penalty. *Magnetic Resonance in Medicine*, *67*(5), 1210–1224. https://doi.org/10.1002/mrm.23097

Smith, R., Keramatian, K., & Christoff, K. (2007). Localizing the rostrolateral prefrontal cortex at the individual level. *NeuroImage*, *36*(4), 1387–1396. https://doi.org/10.1016/j.neuroimage.2007.04.032

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., & Kincses, T. (2019). Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. *NeuroImage*, *185*, 12–26. https://doi.org/10.1016/j.neuroimage.2018.09.078

Taylor, P. A., Reynolds, R. C., Calhoun, V., Gonzalez-Castillo, J., Handwerker, D. A., Bandettini, P. A., Mejia, A. F., & Chen, G. (2023). Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *NeuroImage*, *274*, 120138. https://doi.org/10.1016/j.neuroimage.2023.120138

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, *35*(1), 105–120. https://doi.org/10.1016/j.neuroimage.2006.11.054

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, *8*(1). https://doi.org/10.1214/14-EJS909

Traut, N., Heuer, K., Lemaître, G., Beggiato, A., Germanaud, D., Elmaleh, M., Bethegnies, A., Bonnasse-Gahot, L., Cai, W., Chambon, S., Cliquet, F., Ghriss, A., Guigui, N., De Pierrefeu, A., Wang, M., Zantedeschi, V., Boucaud, A., Van Den Bossche, J., Kegl, B., ... Varoquaux, G. (2022). Insights from an autism imaging biomarker challenge: Promises and threats to biomarker discovery. *NeuroImage*, *255*, 119171. https://doi.org/10.1016/j.neuroimage.2022.119171

Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. *NeuroImage*, *16*(3), 765–780. https://doi.org/10.1006/nimg.2002.1131

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), 62. https://doi.org/10.1038/s42003-018-0073-z

Turner, B. O., Santander, T., Paul, E. J., Barbey, A. K., & Miller, M. B. (2019). Reply to: fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 129. https://doi.org/10.1038/s42003-019-0379-5

Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J., & Coalson, T. (2012). Parcellations and Hemispheric Asymmetries of Human Cerebral Cortex Analyzed on Surface-Based Atlases. *Cerebral Cortex*, *22*(10), 2241–2262. https://doi.org/10.1093/cercor/bhr291

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Varoquaux, G. (2017). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(June 2017), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., . . . Weiss, S. R. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7. https://doi.org/10.1016/j.dcn.2017.10.002

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine*, *368*(15), 1388–1397. https://doi.org/10.1056/NEJMoa1204471

Wager, T. D., Phan, K., Liberzon, I., & Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: A meta-analysis of findings from neuroimaging. *NeuroImage*, *19*(3), 513–531. https://doi.org/10.1016/S1053-8119(03)00078-8

Wang, G., Muschelli, J., & Lindquist, M. A. (2021). Moderated t-tests for group-level fMRI analysis. *NeuroImage*, *237*, 118141. https://doi.org/10.1016/j.neuroimage.2021.118141

Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding Animate Agents: Distinct Roles for the Social Network and Mirror System. *Psychological Science*, *18*(6), 469–474. https://doi.org/10.1111/j.1467-9280.2007.01923.x

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058

Xu, J., Moeller, S., Strupp, J., Auerbach, E., Chen, L., Feinberg, D. A., Ugurbil, K., & Yacoub, E. (2012). Highly accelerated whole brain imaging using aligned-blipped-controlled-aliasing multiband epi. *Proceedings of the 20th Annual Meeting of ISMRM*, *2306*, 1907–1913.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x

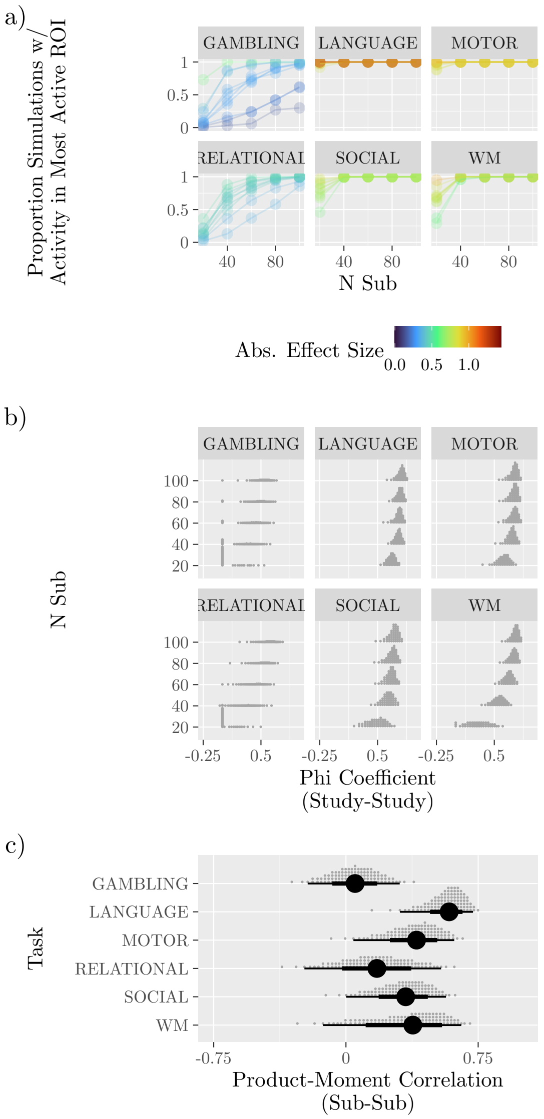

Figure 2: Activation Within Regions of Interest. **a)** Comparison between Gold Standard and Simulations. Effect size refers to the average effect size of voxels within an ROI in the gold standard. **b)** Distribution of Pairwise Mean Square Contingency (Phi) Coefficient for Activations across Simulations. Each dot corresponds to one percentile. **c)** Pairwise Correlation of Effect Size Maps across Subjects. Dots within the distributions correspond to percentiles. The large black dots mark medians and are surrounded by bars marking the 0.66 and 0.95 quantile intervals. All) Regions were defined according to the Schaefer 400 parcellation.

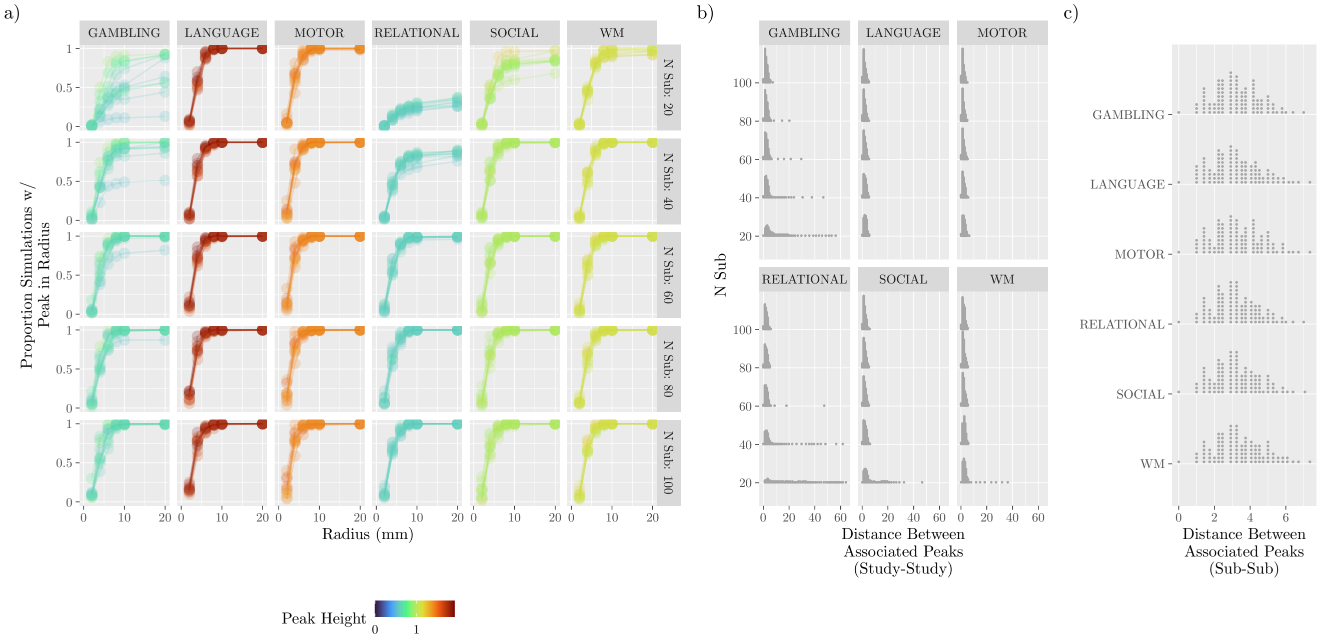

Figure 3: Localization of Peak Activation. **a)**. Proportion of Simulated Studies Containing Peaks within a Given Radii. Columns indicate tasks from the HCP dataset, and rows are the sample sizes that were simulated. Each panel depicts the proportion of simulated studies that contained peaks that were within a given radius of one of the 10 largest peaks for that task, with peaks colored by the effect size of that voxel in the gold standard. **b)** Average Distance Between Peaks in Simulated Studies that were Associated with Common Peaks in the Gold Standard. Within distributions, points represent percentiles of distances. **c)** Average Distance Between Peaks in Individual Subject Maps that were Associated with Common Peaks in the Gold Standard. Within distributions, points represent percentiles of distances.

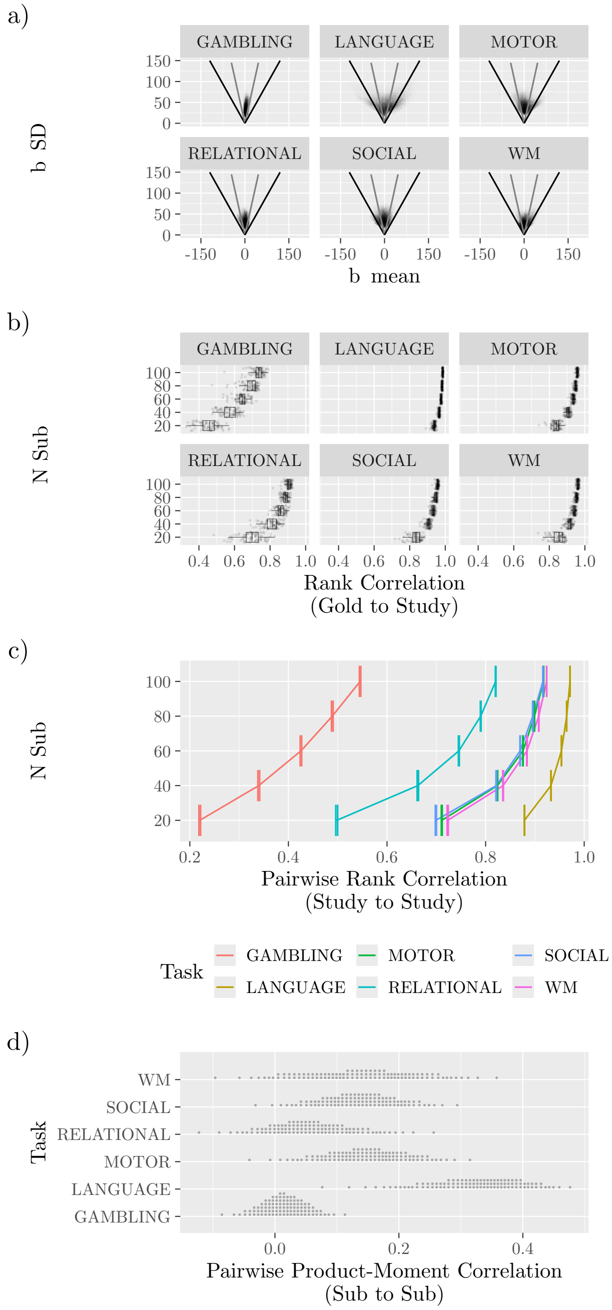

Figure 4: Reliability of Topographic Maps. **a)** Gold Standard. Points indicate voxel-wise standard deviation vs. mean of the contrast estimates. The gray and black lines mark the significance thresholds for a one-sample t-test at $p = 0.001$, for sample sizes $N = 20$ (black) and $N = 100$ (gray) **b)** Comparison between Gold Standard and Simulations. Points mark rank correlations between simulated studies and the gold standard and are summarized with box plots. **c)** Pairwise Rank Correlations Across Simulations. Error bars span $95\,\%$ Confidence Intervals (bootstrapped), and the lines trace averages. **d)** Pairwise Correlation of Voxel-Wise Effect Size Maps across Subjects. Dots within the distributions correspond to percentiles.

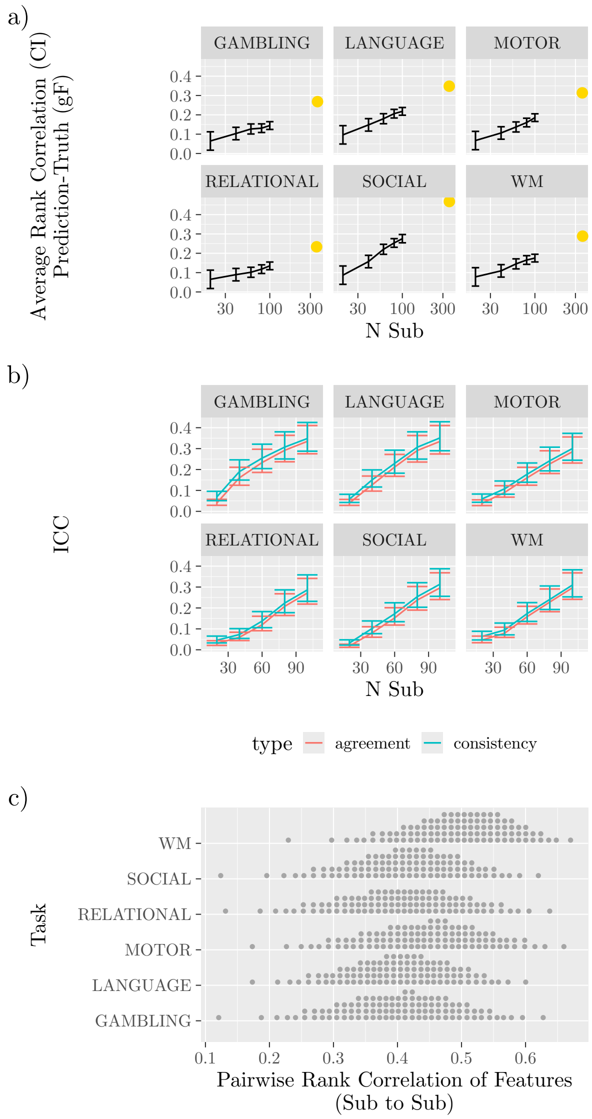

Figure 5: Reliability of Multivariate Pattern Models. **a)** Gold Standard and Comparison between Gold Standard and Simulations. The lines trace the rank correlation between model predictions and the true values in held-out samples (random-effects meta-analysis of correlations across simulated studies), with error bars spanning $95\,\%$ confidence intervals. The isolated points mark gold-standard performance. **b)** Reliability of Model Predictions Across Splits. The intraclass correlation (ICC) was based on a 2-way random effects model. **c)** Pairwise Rank Correlations across Subjects. Correlations were calculated with the model features that were used for training. Dots correspond to percentiles.

1 HCP Tasks

The following task descriptions are paraphrased from the details provided by Barch et al. (2013).

Working Memory

In the working memory runs, subjects (494) completed N-back tasks (Drobyshevsky et al., 2006). Each of the two runs contains eight stimulus blocks consisting of ten trials ($2\,\mathrm{s}$ stimulus presentation, $500\,\mathrm{ms}$ ITI) and 4 fixation blocks ($15\,\mathrm{s}$ each). Within each run, four different stimulus types (faces, places, tools, and body parts) were presented in separate blocks. Task type was indicated at the start of each block with a $2.5\,\mathrm{s}$ cue. Each block contains 2 targets and 2–3 non-target lures. Half of the blocks for each stimulus type used a 2-back task and the other half a 0-back task.

The design matrix included eight task-related predictors, one for each stimulus type in each of the N-back conditions. Each predictor covered the period from the onset of the cue to the offset of the final trial. Our analyses considered a contrast selecting for the 2-back parameter, the 0-back parameter, and a comparison of the 2-back minus 0-back parameters.

Motor

In each of two runs, subjects (492) were asked to move different body parts (Buckner et al., 2011; Thomas Yeo et al., 2011). The body part to move was indicated by visual cues presented at the start of each block (one $3\,\mathrm{s}$ cue for each $12\,\mathrm{s}$ block). Cues indicated that subjects should either tap their left or right fingers, squeeze their left or right toes, or move their tongue (only one type of motion was asked for in each block). Each run contained two blocks of tongue movements, two blocks of each hand movement, two blocks of each foot movement, and three additional blocks of fixation (each $15\,\mathrm{s}$).

The design matrix included five task-related predictors, each covering the duration of the 10 movement trials ($12\,\mathrm{s}$). The cue was modeled separately. Our analyses considered a contrast selecting for the cue, one selecting for the average of motion, and a comparison of average motion minus baseline.

Gambling

In this task, subjects (494) were asked to guess the number on a hidden card in order to win or lose money (Delgado et al., 2000). They were informed that the number ranged from 1–9, and that they should guess whether the hidden number was greater or less than 5. Guesses were realized by pressing one of two buttons. The task is presented in blocks of eight trials that comprised either mostly rewards (6 reward trials interleaved with either 1 neutral and 1 loss trial, 2 neutral trials, or 2 loss trials) or mostly losses (6 loss trials interleaved with either 1 neutral and 1 reward trial, 2 neutral trials, or 2 reward trials). After responding, subjects were given feedback. On reward trials, the feedback was a green up arrow with $1, on loss trials it was a red down arrow with $0.50, and on neutral trials with was the number 5 with a gray, double-headed arrow. In each of the two runs, there were two mostly reward and two mostly loss blocks, interleaved with four fixation blocks (each $15\,\mathrm{s}$).

The design matrix included two task-related predictors that modeled the mostly reward and mostly punishment blocks, each covering the duration of 8 trials ($28\,\mathrm{s}$). Our analyses relied on contrasts for the reward parameter, the punish parameter, and the reward minus punish parameters.

Language

In each run of this task, subjects (483) listened to eight blocks of stimuli, four of which consisted of short stories and four of which consisted of arithmetic (Binder et al., 2011). Blocks averaged $40\,\mathrm{s}$ and the two stimulus types were interleaved. After each block, subjects were presented with a two-alternative forced choice that either asked subjects about the story or the result of the arithmetic.

The design matrix included two predictors that corresponded to the two types of stimuli. Our analyses relied on contrasts for the story parameter, the math parameter, and math minus story.

Relational

In this task, subjects (481) were presented with sets of stimuli and discerned whether they matched or differed according to pre-specified rules (R. Smith et al., 2007). Stimuli varied according to shape and texture. In a relational condition, two pairs of stimuli were presented. One pair serves as a reference, and the stimuli in that pair differ along one of the two dimensions. Subjects first determined the mismatching dimension, and then they determined whether the other pair of stimuli differed along that same dimension. In a matching condition, a single pair of object stimuli was presented along with a third object and a word. The word identified one of the two features, and subjects had to determine whether that identified feature in the third stimulus matched the feature value for either of the paired stimuli. Relational stimuli were presented for $3500\,\mathrm{ms}$ (ITI: $500\,\mathrm{ms}$) and matching stimuli for $2800\,\mathrm{ms}$ (ITI: $400\,\mathrm{ms}$). Stimuli were presented in three blocks that each contained five trials, and the stimulus blocks were interspersed with three fixation blocks $(16\,\mathrm{s})$.

The design matrix for this task included two predictors, one for each of the two conditions (“match” and “relation”). Our analyses relied on contrasts for the matching minus relation parameters.

Social

The social task used stimuli derived from Castelli et al. (2013) and Wheatley et al. (2007). The stimuli were videos of shapes that moved either randomly or with a set of specified interactions $(20\,\mathrm{s}$ per video). After each video, subjects (486) selected one of three responses about whether they observed the objects interacting. There were five blocks of trials per run (conditions balanced across the two runs).

The design matrix for this task contained two predictors, one for each of the “theory of mind“ and “random” conditions. Our analyses used contrasts that selected the subtraction of the theory of mind from the random parameter.

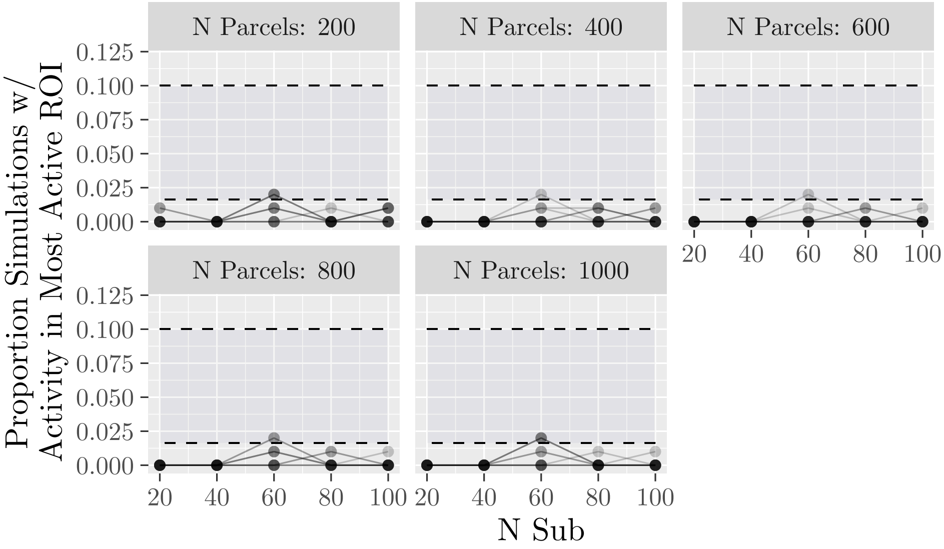

Figure S1: False Positive Rate for Probabilistic Threshold-Free Cluster Enhancement. A "null" dataset was simulated following the methods described by (Wang et al., 2021), in which the Working Memory contrasts were calculated on resting state data. Any positive activation is therefore a false positive. As with the main text, the focus is on the 10 regions exhibiting the higest activation in the gold-standard; regions are grouped by lines. The ribbons span the exact (Clopper-Pearson) 95% Confidence Interval for an error rate of 5%, assuming a binomial distribution with 100 samples (Thulin, 2014).

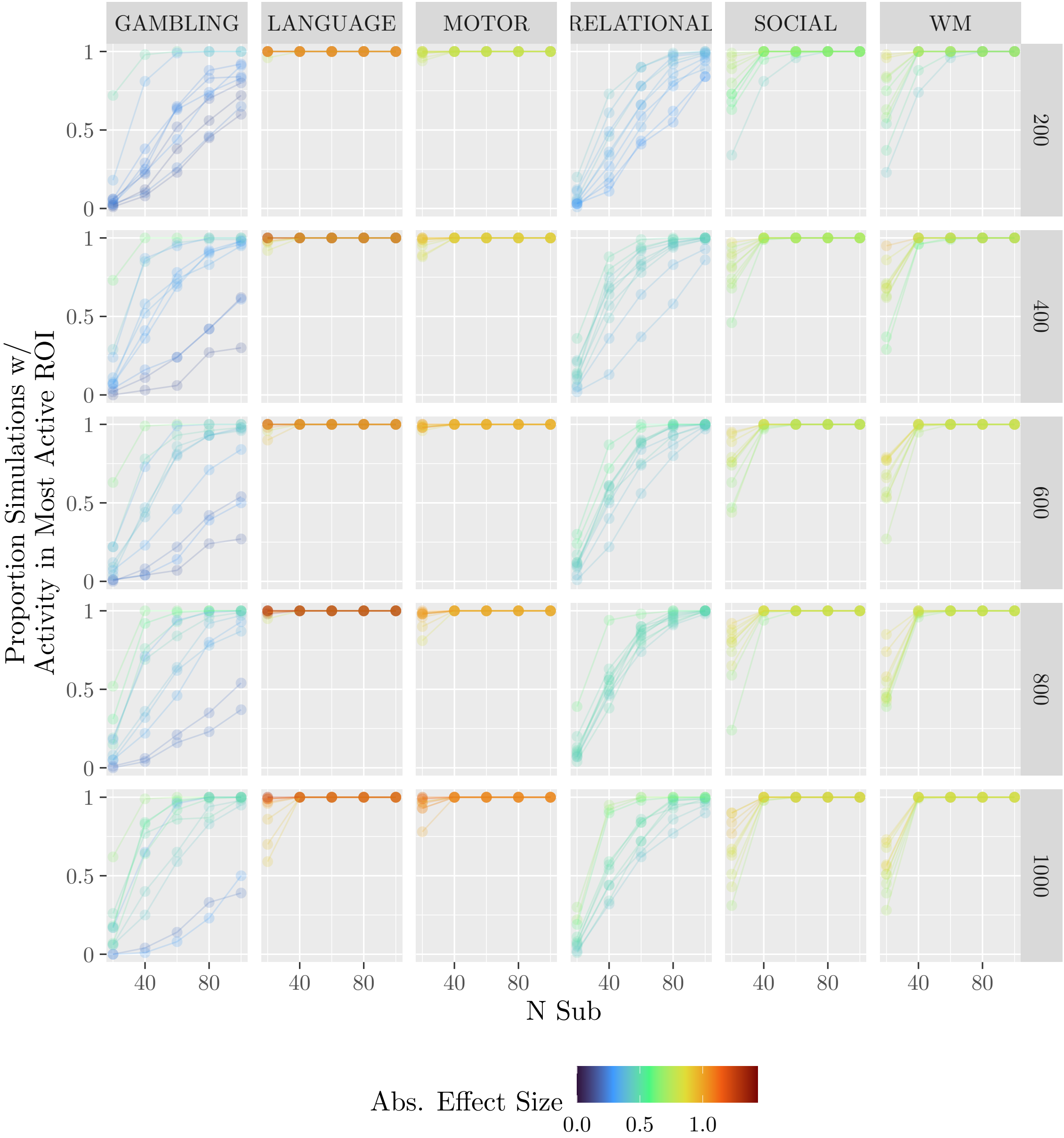

Figure S2: Recovery of Gold Standard across Parcellations. Rows indicate the number of parcels within the atlas. Compare with Figure 2a.

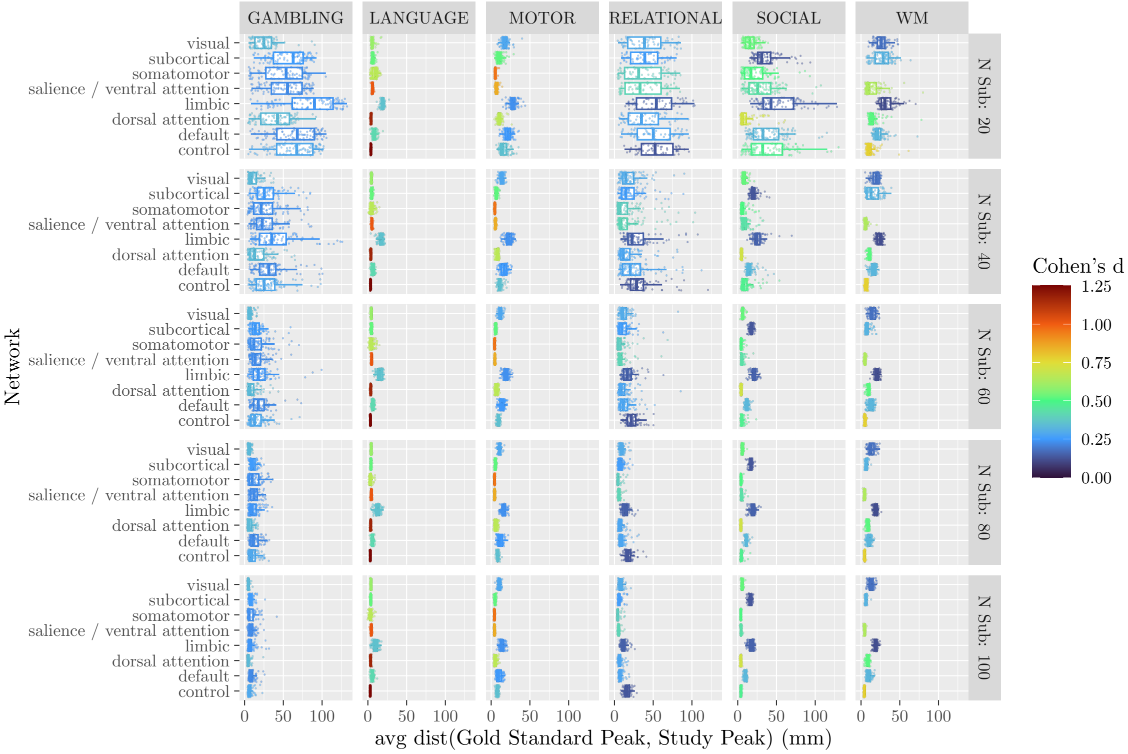

Figure S3: Distribution of Average Distances within Yeo7 Networks. Simulations are grouped by the sample size (rows) and the task (columns). Peaks were selected from each reference map and labeled according to the Yeo7 networks. Across all peaks within a network, the distance to the nearest study peak was calculated, and then these distances were averaged by simulated study, network, task, and sample size. Points mark the distance to the nearest peak within each simulated study. Colors indicate the average effect size of peak voxels in the gold standard for the given network. Note that some panels lack points for some networks because peaks were not identified in those networks in the gold standard.

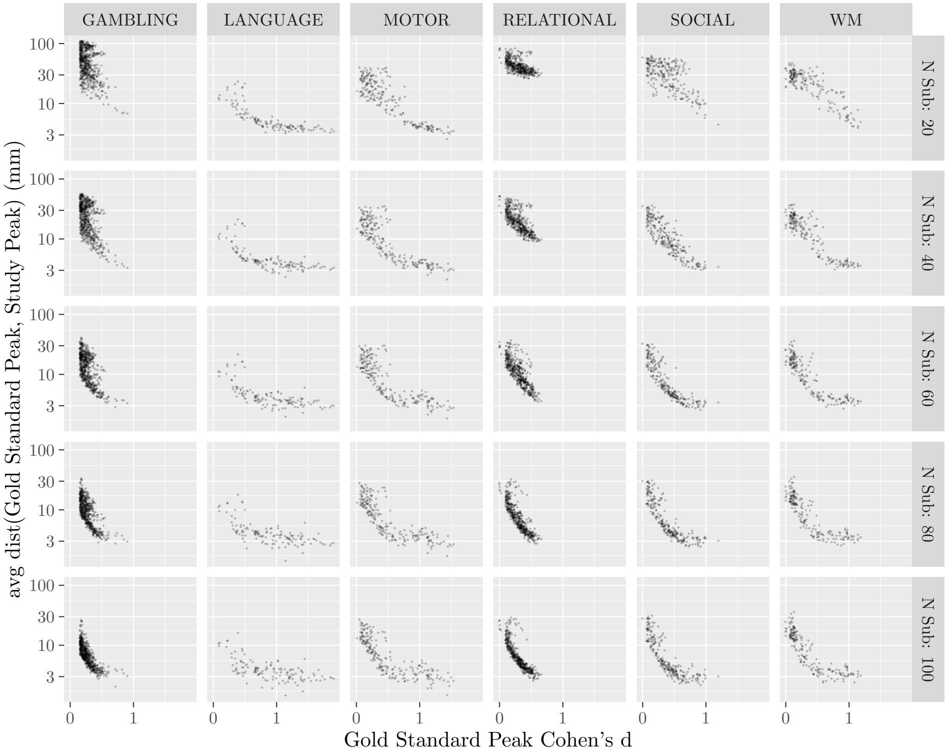

Figure S4: Average Distance from Reference Peak by Reference Effect Size. Individual points correspond to peaks in the reference dataset. Averages were taken across simulated studies and grouped in columns by sample size. Note that the y-axis is on a log scale. Peaks were restricted to those that survived familywise error corrections (0.05).

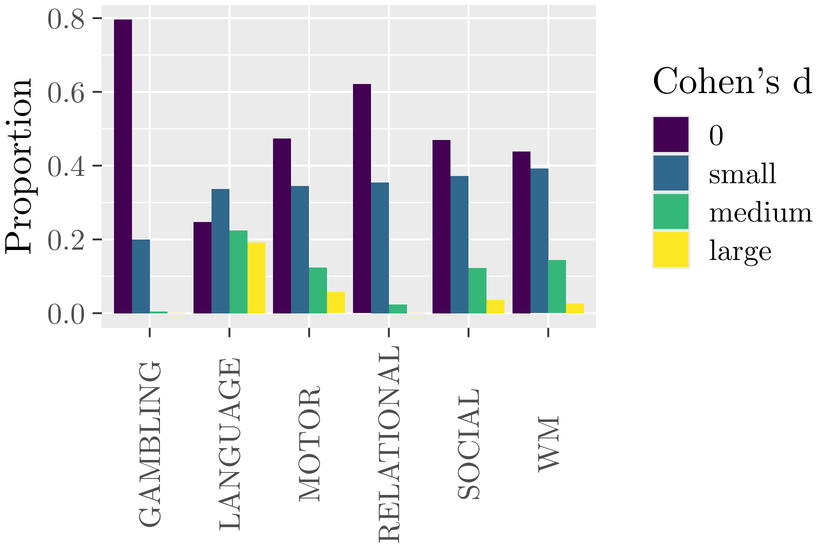

Figure S5: Voxel-wise Effect Sizes Produced by Each Task. Categories were defined as specified in the Methods.

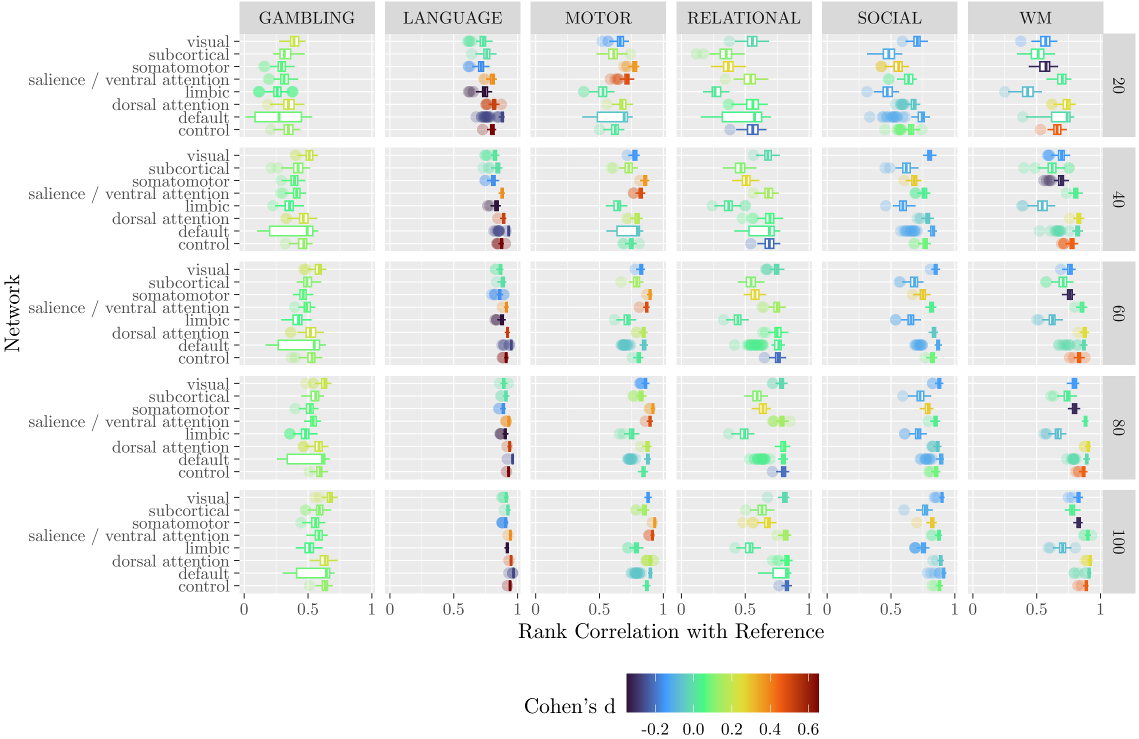

Figure S6: Recovery of Gold Standard Topography by Yeo7 Networks. Colors indicate the average effect size of voxels assigned to the network within the gold standard. Points correspond to simulation iterations.

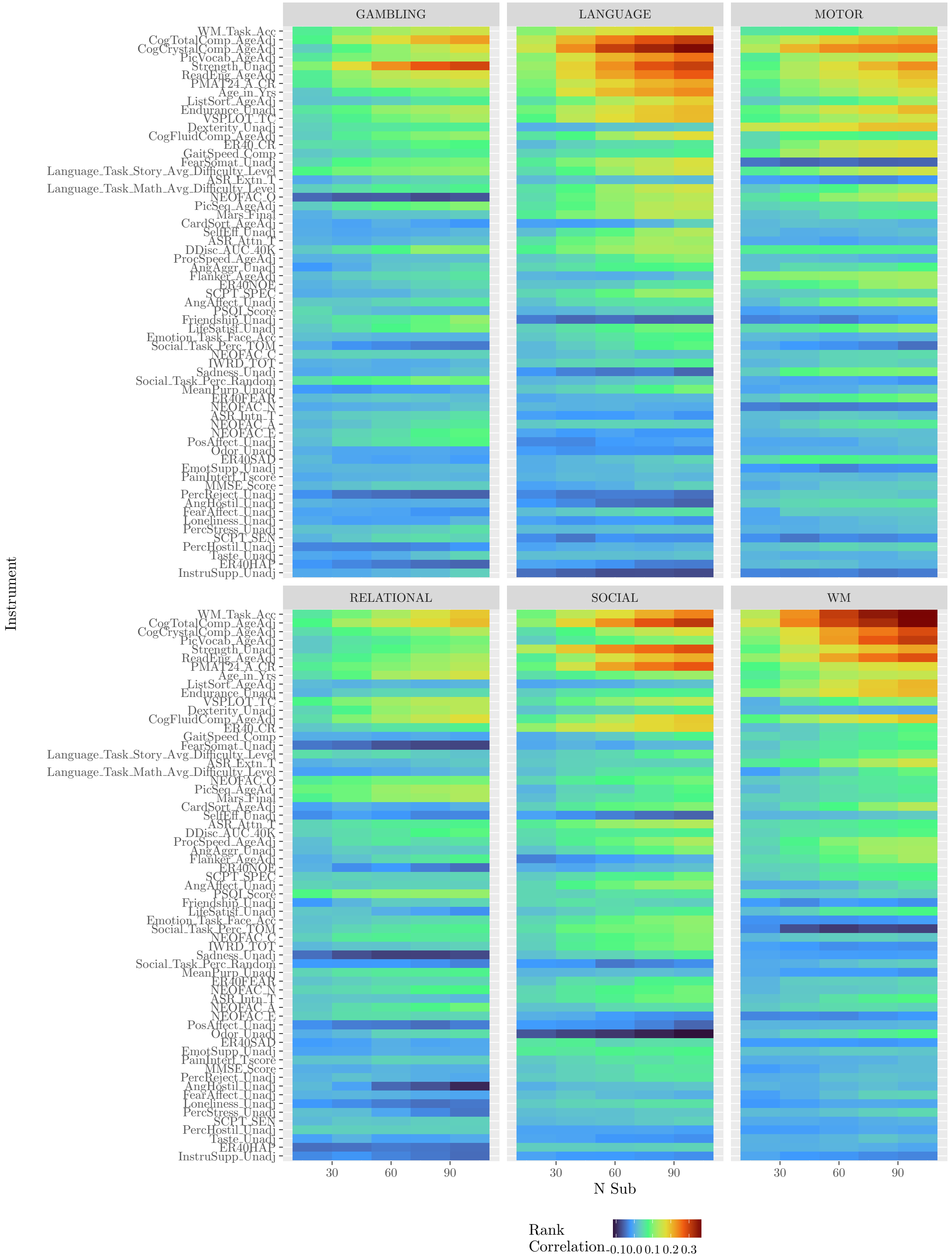

Figure S7: Prediction of Additional Instruments in HCP by Connectivity of DiFuMo components. Color corresponds to rank correlation (measured with meta-analysis across simulation repetitions, as in the main text). Instruments are ordered by maximum estimated rank correlation. The measure of fluid intelligence discussed in the main text corresponds to PMAT24_A_CR.

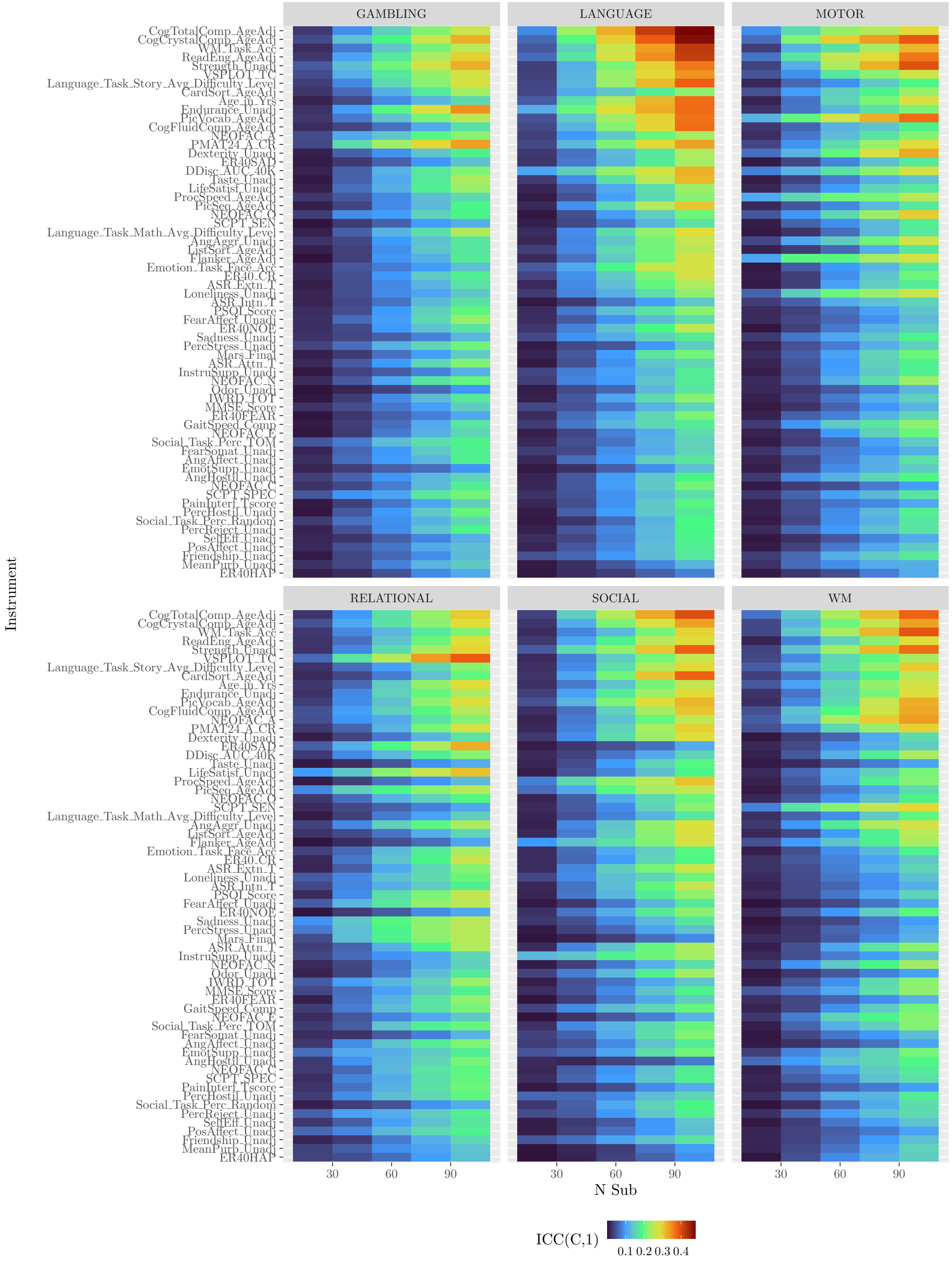

Figure S8: Agreement of Predictions on Test Samples across Simulations. The measure of fluid intelligence discussed in the main text corresponds to PMAT24_A_CR.

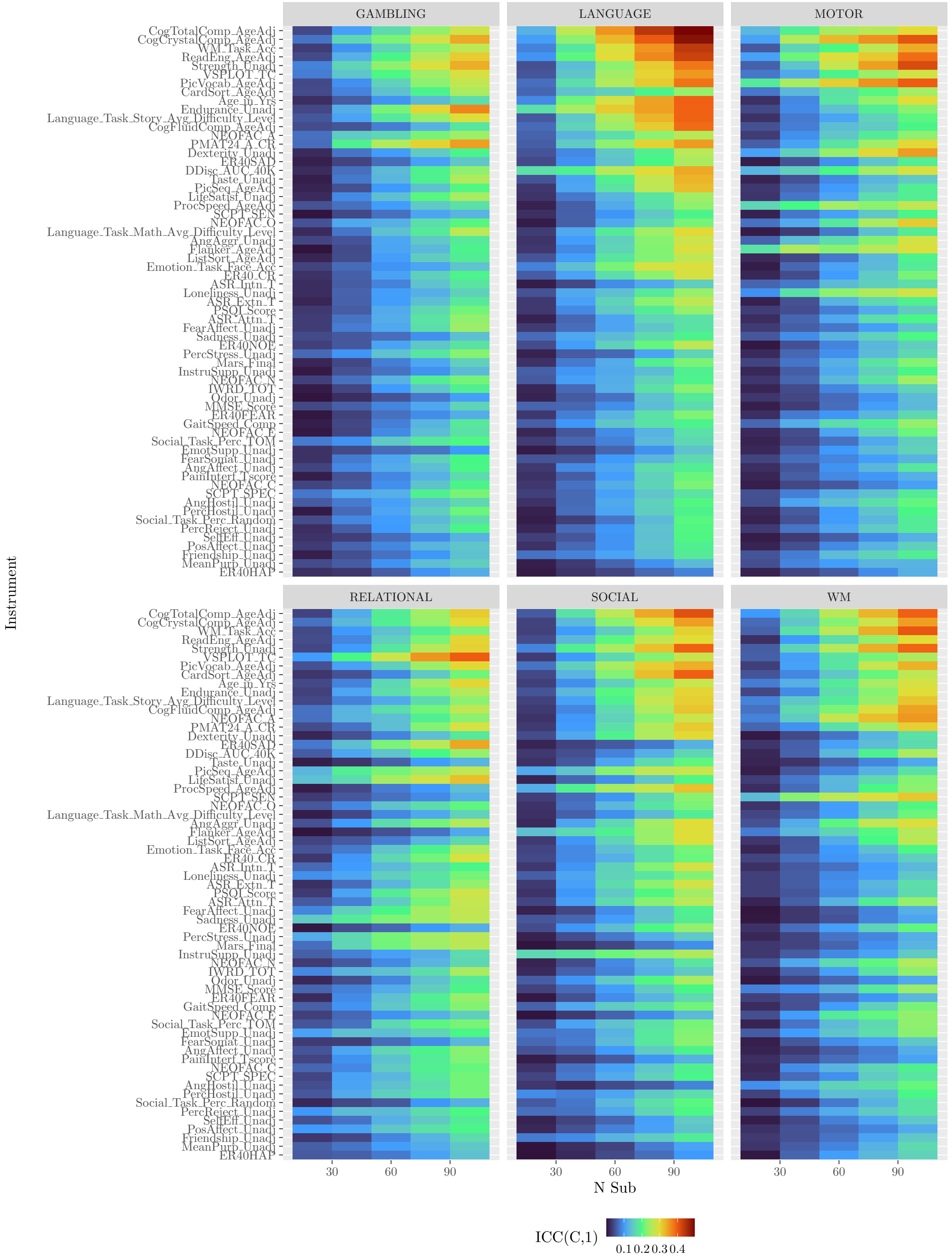

Figure S9: Consistency of Predictions on Test Samples Across Simulations. The measure of fluid intelligence discussed in the main text corresponds to PMAT24_A_CR.