HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2022 February 25.

Published in final edited form as:

Neuroimage. 2021 December 15; 245: 118656. doi:10.1016/j.neuroimage.2021.118656.

The role of neural load effects in predicting individual differences in working memory function

Y. Peeta Li[a,*], **Shelly R. Cooper**[b], **Todd S. Braver**[b]

[a]Department of Psychology, University of Oregon, 1227 University St, Eugene, OR 97403, United States

[b]Department of Psychological and Brain Sciences, Washington University in Saint Louis, 1 Brookings Drive, Saint Louis, MO 63130, United States

Abstract

Studies of working memory (WM) function have tended to adopt either a within-subject approach, focusing on effects of load manipulations, or a between-subjects approach, focusing on individual differences. This dichotomy extends to WM neuroimaging studies, with different neural correlates being identified for within- and between-subjects variation in WM. Here, we examined this issue in a systematic fashion, leveraging the large-sample Human Connectome Project dataset, to conduct a well-powered, whole-brain analysis of the N-back WM task. We first demonstrate the advantages of parcellation schemes for dimension reduction, in terms of load-related effect sizes. This parcel-based approach is then utilized to directly compare the relationship between load-related (within-subject) and behavioral individual differences (between-subject) effects through both correlational and predictive analyses. The results suggest a strong linkage of within-subject and between-subject variation, with larger load-effects linked to stronger brain-behavior correlations. In frontoparietal cortex no hemispheric biases were found towards one type of variation, but the Dorsal Attention Network did exhibit greater sensitivity to between over within-subjects variation, whereas in the Somatomotor network, the reverse pattern was observed. Cross-validated predictive modeling capitalizing on this tight relationship between the two effects indicated greater predictive power for load-activated than load-deactivated parcels, while also demonstrating that load-related effect size can serve as an effective guide to feature (i.e., parcel) selection, in maximizing predictive power while maintaining interpretability. Together, the findings demonstrate an important consistency across within- and between-subjects approaches to identifying the neural substrates of WM, which can be effectively harnessed to develop more powerful predictive models.

*Corresponding author. yichen.li@wustl.edu (Y.P. Li).

Credit authorship contribution statement

Y. Peeta Li: Conceptualization, Visualization, Validation, Writing – original draft, Writing – review & editing. **Shelly R. Cooper:** Conceptualization, Visualization, Validation, Writing – original draft, Writing – review & editing. **Todd S. Braver:** Conceptualization, Visualization, Validation, Writing – original draft, Writing – review & editing.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118656.

Keywords

Working memory; N-back; Load-related effect; Individual difference; Parcellation

1. Introduction

A major line of cognitive neuroscience research has been directed towards understanding the neural basis of working memory (WM). This work has tended to adopt one of two methodological approaches, focusing on either within-subject effects or between-subjects differences (i.e., individual differences; Braver et al., 2010). In the within-subject WM studies, a central goal has been to identify the neural correlates underlying WM load-related effects, finding brain regions sensitive to WM maintenance demands (Feredoes and Postle, 2007; Jha and McCarthy, 2000; Motes and Rypma, 2010; Pessoa et al., 2002; Rypma et al., 1999; Veltman et al., 2003) and/or the effects of parametric variation in WM load (Braver et al., 1997; Lamichhane et al., 2020; Van Snellenberg et al., 2015). In contrast, between-subjects WM studies have focused on understanding the neural basis of individual differences in WM function, which have long-been established as a major component of this domain (Baddeley, 2012; Engle et al., 1999; Just and Carpenter, 1992; Saults and Cowan, 1998). In neuroimaging studies of this type, the goal has been to identify key brain regions, for which activity levels correlate with individual variation in WM behavioral performance (Mitchell and Cusack, 2008; Xu and Chun, 2006). Despite the significant advances that have come from each methodological approach, a key unresolved question is the extent to which within- and between-subjects variations in WM reflect the same or dissociable underlying neural systems (Yarkoni and Braver, 2010). The primary goal of the current study was to resolve the degree to which the neural correlates of within- and between-subjects WM variations overlap.

There are good reasons to think that the neural substrates of within- and between-subjects WM variation to be anatomically and statistically dissociable. By design, brain regions revealed by within-subject analyses (e.g., neural load effects) show the most consistent activation patterns within a selected sample; conversely, regions identified by between-subject analyses will have a large component of variability that correlates with WM performance. Importantly, however, the most consistently activated regions do not necessarily have large variability in activation (Yarkoni and Braver, 2010). This is unsurprising since, statistically, between-subjects variability constitutes the error term in within-subject analyses, creating a potential tradeoff between within- and between-subjects effect sizes, all else being equal. Indeed, although the neuroimaging literature has converged on key brain networks, such as the frontoparietal network, as being critical for WM function, it has also revealed spatial dissociations between regions sensitive to within- and between-subjects WM variation. Early neuroimaging studies highlighted the importance of the prefrontal regions for WM load effects (Cairo et al., 2004; Postle et al., 2001; Rottschy et al., 2012), but other regions, such as the anterior cingulate cortex and parietal regions, were identified based on between-subject variation (Bunge, 2001; Todd and Marois, 2005; Xu and Chun, 2006). A more recent study that utilized data-driven approaches in a large sample dataset also highlighted a similar dissociation (Egli et al., 2018). In particular,

the authors identified a parietally-centered network that was sensitive to load-related individual differences, whereas a frontally-centered network was found to be sensitive to load-independent attention level. Yet the literature is still lacking a systematic, whole-brain focused investigation of this issue. The goal of the current study is to fill this gap.

A recent shift in emphasis within the cognitive neuroscience literature has been an appreciation of the importance of establishing predictive power when examining brain-behavior relationships. In particular, predictive power can only truly be established when examining out-of-sample data, such as through the use of cross-validation approaches (Yarkoni and Westfall, 2017). Previous studies that have examined the predictive power of WM-related brain activity have tended to use whole brain activity patterns in a non-selective manner when constructing models to predict behavioral performance (Pornpattananangkul et al., 2020; Satterthwaite et al., 2013; Sripada et al., 2020). However, It has also been recognized that there is a trade-off between predictive and explanatory power, which can often be optimized by favoring models with fewer selective explanatory variables, as these may have a greater potential for interpretation and generalization (Kampa et al., 2014). Indeed, it is not yet known whether some brain regions may exhibit greater predictive power than others, and whether specific functional properties (e.g., neural load effects) might drive these predictive differences. For example, it is well established that brain regions show different functional properties in response to increasing WM loads, with some regions increasing activity, while others – predominantly in the default mode network – show a deactivation pattern (Buckner et al., 2008; Owen et al., 2005). Both load-activated and load-deactivated regions have been found to predict WM performance (Satterthwaite et al., 2013). Nevertheless, there has yet to be a systematic examination testing whether the magnitude and the direction of load-related effects can serve as an informative guide regarding the predictive power of brain regions. Thus, a secondary goal of the current study was to precisely quantify and compare brain regions in terms of their relative predictive power, using within-subjects WM variation (neural load effects) as an index by which to rank-order brain regions. Moreover, we take advantage of a machine learning approach combined with permutation testing to quantify how predictive power of brain regions changes as a function of load-related effect size and sign (i.e., load-activated vs. load-deactivated).

A final goal of the present study was to demonstrate the utility of a parcel-based approach to whole-brain WM analysis, in contrast to the traditional voxel/vertex-based (+ clustering) approaches that have been deployed in the prior literature. In particular, we argue that the parcellation approach is advantageous, because it achieves the goal of dimensionality reduction in a principled manner (by relying on available pre-specified parcellation atlases, so is unbiased), while potentially increasing WM-related effect sizes. This potential gain in effect size is critical due to two reasons. First, a major goal of WM research is to identify neural correlates in terms of the degree of within-subject or/and between-subject variation they can capture (Yarkoni and Braver, 2010). Second, however, prior work has suggested that voxel/vertex-based analyses yield small effect sizes more generally, such that these approaches may be unreliable and/or have insufficient sensitivity when relying on the standard sample sizes employed in task fMRI research (Poldrack et al., 2017; Rottschy et al., 2012).

To achieve these study goals, we utilized the Human Connectome Project (HCP) 1200-release open dataset (Van Essen et al., 2013), along with two distinct parcellation schemes, to systematically examine the relationship among within- and between-subjects variation in the N-back WM task. The Gordon et al. (2016) parcellation has recently been successfully utilized with this task and dataset to conduct analyses of individual differences in WM function (Etzel et al., 2020). The other parcellation that we employed, and used for primary results reporting, is the more recent Schaefer et al. (2018) scheme. The Schaefer parcellation has the advantage of more homogenous parcel sizes and availability in a variety of spatial resolutions (100–1000 parcels). For each scheme, we conducted four sets of analyses: (1) estimating the within-subject effect size of each parcel within each functional network and comparing these parcels to voxel/vertex-level analyses conducted in the same regions (c.f. Poldrack et al. 2017); (2) comparing within- and between-subject effect size, in terms of their relative magnitude and consistency; (3) examining how predictive power changes as a function of the sign and magnitude of the load-related effect size; and (4) testing whether load-related effect size can be utilized as a key indicator variable to guide feature selection, when building predictive models of WM performance.

2. Methods

2.1. Data collection and preprocessing

Neuroimaging data acquired from fMRI scans performed on 1083 healthy adults, in the age range of 22–35 years, were made available through the Human Connectome Project (HCP). The 1200 subject release dataset was used for this study (http://www.humanconnectomeproject.org/). Participants were recruited from the area surrounding Washington University in St. Louis (St. Louis, MO). All participants were given extensive telephone screening interviews and signed the informed consent document at the beginning of the study (see Van Essen et al. (2013) for more detailed information regarding the informed consent process and screening interviews).

All functional images were acquired on a 3T Siemens Skyra scanner with a 32-channel head coil (TR = 720 ms, TE = 33 ms, flip angle = 52°, FOV = 208 mm × 180 mm, matrix size = 104 × 90, 72 slices, 2 mm isotropic voxels). More detailed information regarding pulse sequence and data acquisition is provided in previous publications describing the HCP dataset (Uğurbil et al., 2013; Van Essen et al., 2013). The data were collected over a two-day period. The N-back task was used to assess WM function, based on data acquired in the first fMRI session. Out-of-scanner behavioral assessments of both WM function and general cognitive ability were acquired as part of the testing protocol, with measures from the NIH-toolbox acquired on day one and additional non-Toolbox measures acquired on day two (Barch et al., 2013; Van Essen et al., 2013). Preprocessing of fMRI data was implemented using the HCP minimally preprocessed pipeline, which outputs data in CIFTI format on the associated grayordinates spatial coordinate system. Procedures for the HCP pipeline have been comprehensively described in previous publications (Glasser et al., 2013; Van Essen et al., 2013).

2.2. Dimension reduction using predefined parcellation schemes

For the tasks included in the HCP dataset, 3D spatial maps of the contrast of parameter estimates (COPEs) were computed with FSL software (Smith et al., 2004). These COPEs were released as part of the HCP publicly available distribution package, and reflect the magnitude of brain activation differences between task conditions (i.e., within-subject effects). For the current study, we used COPE #11 in the HCP N-back task, which provides the contrast estimate between the high-load condition and the low-load condition (i.e., 2-back – 0-back), averaging across different stimulus types. Thus, a positive parameter estimate indicates higher activation in the high-load compared to low-load condition (which we hereafter refer to as load-activated), whereas a negative parameter estimate indicates the reverse pattern (which we hereafter refer to as load-deactivated). The COPE data were then summarized, by averaging vertex-wise estimates into predefined parcels. We used two independent, predefined parcellation schemes for dimension reduction in order to assure that our findings are not the result of a particular parcellation scheme but are generalizable. We used the 400 cortical parcels (7 networks) atlas provided by Schaefer et al. (2018) and the 333 cortical parcels (13 networks) atlas provided by Gordon et al. (2016). Although the Schaefer parcellation has various levels of resolution (100–1000 parcels), the 400-parcel set was selected here because it was the one most thoroughly examined in the original paper (Schaefer et al., 2018), and it is close in size, and thus comparable, to the Gordon parcellation. We report all results for the Schaefer 400 parcellation in the main text; Gordon parcellation results are included in the Supplementary Materials, as they were very similar in all respects.

2.3. Behavioral measures

The N-back task was the primary in-scanner task used to assess WM function in the HCP. The N-back is probably the most popular neuroimaging paradigm for assessing WM function, via load (and content) manipulations (Barch et al., 2013; Braver et al., 1997; Lamichhane et al., 2020; Owen et al., 2005). The task included two runs of four blocks, which consisted of four distinct visual-spatial stimulus types, including places, tools, faces and body parts. In the analyses presented here, we ignored the manipulation of stimulus type, and collapsed the data across these four conditions. Although examinations of WM content manipulations are also an important focus of investigation, it was beyond the scope of the current study; consequently, the decision to collapse across stimulus type was made to increase the statistical power and reliability of the behavioral performance estimates (and COPEs). Specifically, the in-scanner behavioral variable of interest was accuracy in the 2-back condition. For the 2-back condition, participants were asked to decide whether the stimulus presented on the current trial was the same as the stimulus two trials back. The relationship between this in-scanner working memory performance measure and the COPE parameter estimate was used to compute the individual difference (i.e., between-subjects) effect size, which was compared with the WM load (i.e., within-subject) effect size in an equivalent manner.

As part of the HCP protocol, participants performed several cognitive tasks in an out-of-scanner behavioral session. Here we selected four out-of-scanner behavioral measures to test the degree to which load-related N-back activity provides a more generalizable

indicator of individual differences in WM capacity and cognitive functioning, by predicting out-of-scanner indices. The rationale is that the out-of-scanner measures should be less impacted by any covarying state-related or non-specific factors that might be reflected in N-back performance. Specifically, we selected the List Sorting task which examines WM capacity; the Picture Vocabulary which examines Language/vocabulary comprehension; Oral Reading Recognition which examines language and reading decoding; and the Penn Matrix Reasoning task which examines general fluid intelligence (Barch et al., 2013). These out-of-scanner tasks were selected because they either directly probe WM capacity (Tulsky et al., 2014) or have been found to be highly related to individual differences in WM function in prior work (Cooper et al., 2019; Pornpattananangkul et al., 2020). In the current study, the primary focus was on the List Sorting task, as it is an explicit measure of WM capacity. Since the other out-of-scanner measures do not directly tap into WM capacity *per se*, these analyses were primarily conducted for comparison, benchmarking, and generalization purposes, and are reported in Supplementary Materials.

2.4. Outlier exclusion

We identified outlier parcels based on extreme COPE values, using a cutoff of above or below 3 times their interquartile range (known as the 3 IQR rule; cf., Pornpattananangkul et al. 2020). Participants with 10% or more of their parcels showing extreme values were identified as outliers and excluded from further analyses. This outlier detection approach was intended to remove participants with potentially poor brain registration. The 3 IQR rule was also applied to the behavioral measures and did not identify any outliers. This criterion excluded 52 outliers and the final sample used in all subsequent analyses included in a total of 989 participants that completed both the in- and out-of-scanner tasks of interest. We also replicated all analyses without any outlier exclusion; all primary results remained unchanged.

2.5. Estimating vertex- and parcel- based measures of load-related effect size

To replicate the voxel-level N-back effect size measured reported in Poldrack et al. (2017), we quantified the load-related effect size of each vertex using Cohen's *d*, which was computed as the mean effect divided by the standard deviation of the data. To provide a network-level estimate of load-related effect size, we grouped all vertices into either 7 (Schaefer) or 13 (Gordon) networks and computed the mean effect sizes within each network. To quantify parcel-level N-back load-related effect size, we first averaged vertex-wise estimates into predefined parcels, and computed parcel-level Cohen's *d* the same way as described above. Parcels were then grouped into either 7 (Schaefer) or 13 (Gordon) networks and the mean parcel-level effect sizes were computed within each network. These parcel-wise effect sizes were then compared with those computed vertex-wise.

2.6. Defining the neural correlates of within- and between-subject variation

For each parcellation scheme, WM within-subject variation was measured in terms of the N-back load-related effect size. We defined neural correlates of the load-related effect by selecting parcels that had significant differences in activation between the high working memory load (2-back) and low load (0-back) conditions (Rottschy et al., 2012). Specifically, we conducted one-sample *t* -tests using the contrast COPE (2-back – 0-back estimates)

against a population mean of 0 to identify parcels exhibiting positive (i.e., load-activated parcels) and negative (i.e., load-deactivated parcels) differences, using a whole brain Bonferroni correction ($p < 0.000125$) to determine statistical significance. Both sets of parcels were load-sensitive and investigated in subsequent analyses involving within-subject WM load effects.

We defined neural correlates of between-subjects WM effects by selecting parcels in which the between-subjects variability in load-related activation was associated with variability in 2-back behavioral performance. Specifically, we computed the Pearson correlation coefficient (r) between each parcel's load-related activation level and the 2-back task performance to identify parcels exhibiting either a positive or negative linear relationship with behavioral performance, again using a whole-brain Bonferroni correction ($p < 0.000125$) to determine statistical significance. Both sets of parcels were investigated in subsequent analyses involving between-subject WM load effects.

2.7. Comparison of within-subject and between-subject WM effects

In order to directly compare within- and between-subject WM effects, we computed a normalized ranking of both effect sizes across the full set of parcels. The Cohen's d coefficient was used to quantify and rank order parcels according to their sensitivity to WM load; the Pearson correlation coefficient (r) was used to quantify and rank order parcels according to their sensitivity to individual differences in behavioral performance. Thus, each parcel was assigned two rank scores based on the absolute value of their effect size for within- and between-subjects effects, respectively. Next the Spearman's correlation (r_s) between the two rank scores was computed separately for each of three parcel sets: load-activated, load-deactivated, and load-insensitive (i.e. parcels that did not show a statistically significant load-related effect). A strong positive correlation coefficient indicates that the parcels contribute to within- and between-subjects WM variations in a consistent fashion. Conversely, a weak correlation coefficient indicates that there not a strong linkage between the two types of WM effect.

Further, to visualize the spatial distribution of parcels exhibiting a bias towards within- or between-subjects effect, we computed the difference of the rank scores for each parcel (between - within effect size ranking). Based on this rank difference score, a positive value indicates that a parcel is biased to be more sensitive to within-subject variation, while a negative value indicates that the parcel is biased to be more sensitive to between-subjects variation. These rank difference scores were then normalized and visualized on the brain surface to identify any potential spatial or anatomical gradients in these biases.

2.8. Building univariate predictive models of WM load-related effects

We investigated the degree to which the WM load-related effect size and direction of effect (i.e., load-activated vs. load-deactivated) indicated the predictive power of a parcel. Similar to the correlational approach, predictive models were used to index a parcel's sensitivity to between-subjects variation but from a distinct inferential perspective. In standard correlational analyses, the Pearson's r value is an index that attempts to *explain* behavioral individual differences (i.e., explanatory power) observed in the N-back task, in

terms of load-related activation. Yet such approaches are not explicitly implemented to *predict* behavioral performance of a new individual (i.e., predictive power), based on their neural activity pattern (or to *predict* the behavioral performance of that same individual in a different WM task; Yarkoni and Westfall, 2017). In order to build truly predictive models, it is necessary to utilize cross-validation approaches in which predictions are evaluated on held out (out-of-sample) data.

To provide a benchmark, univariate models were first used to establish the predictive accuracy of each parcel in isolation, estimated through cross-validation. Specifically, for each parcel, a simple linear regression model was trained on 9 folds of the data with the load-related neural activity used to predict individual differences in behavioral performances, with predictive accuracy tested on the left-out fold. The predictive power was quantified as the Pearson correlation coefficient between the predicted and actual behavioral performance, averaged across 10 folds. Next, we ranked each parcel according to its load-related effect size, with separate grouping for load-activated and load-deactivated parcels, to explore how univariate predictive accuracy varied as a function of load-related effect size.

2.9. Building multivariate predictive models of WM load-related effects

A second phase of predictive modeling tested whether multivariate predictive models would outperform univariate models. In particular, prior findings have suggested that the pattern of neural activity across parcels may contain additional information that can be leveraged to increase predictive power (Marek et al., 2020; Pornpattananangkul et al., 2020). To examine this issue systematically, we examined whether the predictive power varied for parcels within different load-related effect size ranges. We took advantage of a machine learning multivariate approach to build predictive models, in which the load-related activation of sets of parcels were used to predict both in- and out-of-scanner behavioral performance in a multivariate manner.

To conduct these types of predictive analyses, we first grouped parcels according to their load-related effect sizes into 12 bins that each spanned a range of 0.2 effect size: 4 bins of load-deactivated parcels with load-related effect sizes ranging from −0.1 to −0.9; 7 bins of load-activated parcels with load-effect sizes ranging from 0.1 to 1.5; and 1 bin of load-insensitive parcels. We included additional load-activated parcels ($N = 8$) in the 1.3 to 1.5 bin and additional load-deactivated parcels ($N = 9$) in the −0.7 to −0.9 bin. For each of the 100 iterations of sampling, we randomly and repeatedly sampled 10 parcels without replacement from each bin, measuring the predictive power at each bin for that sampling; prediction power was then averaged across iteration for each bin. An important benefit of this type of sampling approach is that it enables a comparison of averaged predictive accuracies across bins that is not confounded by the number of predictive features (i.e., number of parcels) in the respective bins. For example, there are 25 load-activated parcels in the 0.1 to 0.3 bin but 58 load-deactivated parcels in the −0.1 to −0.3 bin. As a result, without controlling the number of parcels, it is unclear whether any predictive accuracy differences observed between the two bins were due to the properties of the parcels or the number of predictive features. Conversely, with this analytic approach, we could systematically test

whether changes in predictive power occurred reliably both as a function of the direction (i.e., load-activated versus load-deactivated) and magnitude of load-related effect size.

For each round of sampling, we used support vector regression (SVR) to test the predictive power of each parcel bin through a 10-fold cross-validation framework. For each bin, a linear SVR model (C = 1.0, epsilon = 0.1) was trained on 9 folds of the data and tested on the left-out fold. The predictive accuracy was measured as the correlation between the predicted and the actual performance scores (Satterthwaite et al., 2013; Sripada et al., 2020). The final predictive accuracy for each bin was averaged across 10 folds and across 100 rounds of sampling. Moreover, we also tested the predictive power of each bin on out-of-scanner tasks (e.g., List Sorting) to test whether the inferences derived from these parcel groups generalized to working memory performance more broadly, rather than just in-scanner N-back performance specifically.

In order to statistically quantify how the predictive power of parcels changes as a function of load-related effect size, we developed a unique, nested permutation test, adapted from the permutation paradigm used in Etzel and Braver (2013). As shown in Fig. S1, for each round of sampling, we pooled and shuffled the sampled parcels across the 12 bins (nParcels = 120), randomly assigning each bin a new set of 10 parcels. The same SVR cross-validation framework was applied to measure the null predictive accuracy for each bin. This shuffling process was iterated 1000 times, resulting in a total 1000 null predictive accuracy measures per bin per sampling process. These null measures were then averaged across the 100 sampling processes. Using these null measures, we constructed null distributions for: 1) the beta values for linear trend tests that focused on the effect of bin; and 2) the predictive power differences between any two parcel bins. Thus, we could estimate the probability of observing the linear trend and predictive power differences measured by the real, unshuffled data.

2.10. Using load-related effect size for feature selection

In a last phase of analysis, we explored whether the load-related effect size could be treated as a useful indicator variable from which to select features in building predictive models. In particular, selecting features according to load-related effect size (larger to smaller) could be a useful heuristic that enables a more parsimonious predictive model, i.e., one that provides an optimal mixture of interpretability and explanatory power, combined with maximal predictive accuracy. Specifically, beginning with the parcel with the largest load-related effect size, we sequentially added parcels as features to predict both in-scanner and out-of-scanner working memory performance, using the load-related effect size as the metric by which to add each new feature (parcel). In other words, we iteratively built a set of models, in which each was constructed by successively adding features (parcels) according to rank-ordered effect size, then measuring the change in predictive power as the model accumulated each new feature. The goal of this analysis was to determine if a load-related effect size cutoff could be identified, whereby adding further features (parcels) would no longer improve model performance.

To evaluate this feature selection principle, we compared the observed model performance to predictive models in which the features were randomly selected (i.e., without reference

to load-related effect sizes). Specifically, for each predictive feature size, we randomly sampled without replacement the same number of predictive features and estimated the model performance. This process was performed with 1000 iterations, generating a null distribution for each predictive feature size. Note that due to computational constraints, we only conducted this analysis for predictive models up to 60 features. The 5% to 95% envelope of each null distribution was computed and plotted to benchmark the observed model performance.

2.11. Data and code availability

Behavioral and processed fMRI data supporting the primary findings of this study are available at the 1200s HCP release (http://www.humanconnectomeproject.org/). The code for performing the specific analyses described in this paper can be found through the Open Science Framework at https://osf.io/atkum/.

3. Results

3.1. Parcels have larger effect sizes

The load-related effect size was computed for each vertex (N = 64,984 vertices) or parcel (N = 400 parcels) within each Schaefer network (N = 7 networks; Fig. 1a). We quantified the standardized effect size using Cohen's d. As predicted, parcels located within the frontoparietal control network (FPN; termed "Control" in the Schaefer scheme) showed the highest averaged effect sizes (d = 0.93 ± 0.47; Fig. 1b), with 66% (34/52) of the parcels reaching the level of d = 0.8, which is standardly defined as a large effect size; conversely, only 19% (10/52) of the parcels had less than a medium effect size (d < 0.5). On the other hand, the vertex-level effect sizes within the same network were much lower on average (d = 0.57 ± 0.35; Fig. 1c), which is consistent with the effect size measures reported in Poldrack et al. (2017). Specifically, only 28% (1934/6907) of the vertices in the Control network showed a large effect size (d > 0.8), whereas about 43% (2959/6907) of the vertices had less than a medium effect size (d < 0.5). These results were replicated using Gordon parcellation scheme (Fig. S2a), with FPN and the Dorsal Attention Network (DAN) showing mean parcel-level effect sizes of d = 0.92 ± 0.49 and d = 0.80 ± 0.47 (Fig. S2b), but much smaller mean vertex-level effect sizes of d = 0.59 ± 0.33 and d = 0.44 ± 0.37, respectively (Fig. S2c). This increase of effect size highlights the advantages of utilizing predefined parcellation schemes and parcels as the primary units of analysis, rather than the traditional voxel-wise approach. Importantly, the gain in sensitivity from using parcels did not come at a cost of specificity, as a clear differentiation and interpretable ordering was observed across brain networks. In particular, although large effect sizes were observed in both the FPN/CONT and DAN, other brain networks that are thought to be only weakly associated with WM, such as the Visual and Limbic networks, exhibited small effect sizes centered near zero (and this was the same as was found with vertex-based effect sizes).

3.2. Identification of WM-involved networks

Utilizing the parcel-based approach, the goal of the next set of analyses was to identify the parcels that exhibited statistically reliable within- or between-subjects WM effects. Specifically, parcels showing within-subject WM effect should exhibit consistent activation

changes across all participants in response to increasing WM load; on the other hand, parcels showing between-subjects WM effect should exhibit associations between their load-induced activity variations and behavioral performances. To first identify parcels that showed significant within-subject WM effects, we performed parcel-wise (N = 400 parcels) one-sample t-tests with the preprocessed contrast estimates between 2-back and 0-back (2-back – 0-back) conditions for each subject (N = 989). In total, 353 parcels showed sensitivity to changes in working-memory load demands (whole-brain Bonferroni corrected at $p < 0.000125$; Fig. 2a, Fig. S3a). In particular, among load-sensitive parcels, many exhibited increased activity with increased load (i.e., load-activated parcels; N = 157; Mean Cohen's $d = 0.74 \pm 0.41$; range: 0.12 to 1.75). However, a substantial subset of cortical parcels responded in the reverse manner, with decreased activity associated with increased load (i.e., load-deactivated parcels; N = 196; Mean Cohen's $d = -0.49 \pm 0.24$; range: −1.16 to −0.12). We then used brain-behavior correlations to identify parcels that exhibited between-subjects WM effects. The analysis revealed 177 parcels, for which load-related activity was associated with between-subject differences in 2-back task performance (whole brain Bonferroni correlated $p < 0.000125$; Fig. 2b, Fig. S3b). In this case, most parcels exhibited a positive correlation between load-related activity and 2-back performance (N = 120; Mean $r = 0.24 \pm 0.08$; range: 0.12 to 0.44); a smaller number exhibited a negative correlation (N = 57; Mean $r = -0.17 \pm 0.04$; range: −0.28 to −0.12).

We compared these two types of effects by examining their overlap, via conjunction analysis. Indeed, a high degree of overlap was observed. Almost 99% of the parcels exhibiting between-subjects WM effects (i.e., brain-behavior correlations) also exhibited significant within-subject (i.e., load-related) effects (175/177 parcels). Furthermore, the sign of the two effects was also highly consistent. That is, parcels that exhibited positive correlations with behavioral performance also tended to show load-related increase in activity (94% of the parcels that exhibited positive behavioral correlations were also load-activated; 113/120 parcels) and vice versa (100% percent of the parcels that exhibited negative behavioral correlations were also load-deactivated; 57/57 parcels). This systematic whole-brain analysis of the relationship between neural correlates of the two types of WM effects (within-subjects vs. between-subject) suggests that they are in fact strongly overlapping – rather than discrepant, as might have been assumed on purely statistical grounds – in terms of their spatial distribution.

3.3. Parcels contribute to between and within-subject variations equivalently

To probe this relationship at a finer grain, a correlational approach was used to examine the degree to which the two types of WM effects were coupled. For this analysis, we normalized the effect size of each measure (load-related effect, brain-behavior correlation) by rank ordering each across parcels, and then conducting a Spearman's correlation (r_s) on the two ranks. Fig. 3a illustrates these relationships, for load-activated, load-deactivated, and load-insensitive parcels, respectively. The results show that the two effects were very strongly correlated when considering load-activated parcels ($r_s = 0.79$). However, the effect was significantly weaker for load-deactivated parcels ($r_s = 0.47$; $z_{diff} = -5.09$, $p < 0.001$), suggest that for load-deactivated parcels the coupling was weaker. Moreover, when examining the load-insensitive parcels, there was no relationship between the two effects, as expected (r_s

= 0.11, $p = 0.46$). These results suggest that WM-related neural activity reflects within- and between-subjects variation in a largely consistent fashion, particularly for load-activated parcels.

Nevertheless, it is possible that subtle anatomical dissociations or gradients might be present which favor one type of effect over the other, as has been alluded to in many previous studies (Yarkoni and Braver, 2010). To examine this issue, we visualized differences in the rankings of the two effect-sizes by plotting them on the brain surface, according to the magnitude of difference or bias (i.e., favoring the within-subject or between-subject effect). Specifically, a positive ranking difference would suggest that the within-subject WM effect size of a given parcel is ranked higher than its between-subject WM effect size, thus showing a bias towards within-subject WM variation; conversely, a negative ranking difference would suggest a bias towards between-subject WM variation. Visual inspection of these patterns (Fig. 3b) seems to indicate that parcels which show a bias towards one type of effect are evenly distributed across the cortex, and without a strong pattern of spatial clustering (e.g., prefrontal vs. parietal, left vs. right hemisphere, etc.). The same pattern of results was observed using Gordon parcellation scheme (Fig. S4). To quantify these observations, we counted the number of parcels in the frontal and parietal region per hemisphere that showed bias towards either type of effect. Then Chi-square tests were used to test for the presences of bias. Indeed, we did not observe any spatial distribution biases toward either type of effect across frontal and parietal lobes or left and right hemispheres (all $p > 0.05$; Table 1). Additionally, we tested whether within- and between-subject effect biased parcels were differentially distributed among functional networks. Interestingly, as shown in Table 2, we found that the parcels in the dorsal-attention network were significantly biased toward between-subject variation ($X^2_{1,\ N=46} = 28.17$, $p < 0.001$) whereas parcels in the somatomotor network were significantly biased toward within-subject variation ($X^2_{1,\ N=77} = 28.69$, $p < 0.001$). The same pattern of results was replicated using Gordon parcellation (Table S1). These results suggest that individual differences in working memory are more likely to be contributed by networks associated with higher cognitive function rather than perceptual/sensorimotor function.

3.4. Load-effect sizes indicates parcel's univariate predictive power

Although a tight relationship was observed between within-subject and between-subject WM effects, to quantify whether these effects can truly be considered predictive in nature, a cross-validation approach is required (Yarkoni and Westfall, 2017). In particular, we examined the impact of load-related effect size on the predictive accuracy of each parcel separately, by focusing on out-of-sample test data. The predictive accuracy was quantified as the correlation strength between predicted and actual behavior performance, examined in the out-of-sample data. The results suggested that load-related effect size is highly correlated with a parcel's univariate predictive power (load-activated parcels: $r = 0.90$, $p < 0.001$; load-deactivated parcels: $r = 0.58$, $p < 0.001$). Fig. 4a shows the top and bottom 30 load-activated parcels ranked by load-related effect sizes, and the impact of the load-related effect size in predicting between-subject variation in WM performance is very easily seen. Specifically, the load-activated parcels with the largest effect sizes ($\bar{d} = 1.37$) tended to be the ones showing the strongest predictive power (mean $\bar{r} = 0.32$), whereas those with smaller effect

sizes ($\bar{d} = 0.22$) tended to show weaker predictive power (mean $\bar{r} = 0.10$). A similar pattern was observed for load-deactivated parcels, yet with less prominent effects (Fig. 4b): parcels with the largest effect sizes ($\bar{d} = -1.12$) exhibited the strongest predictive power (mean $\bar{r} = 0.19$), while parcels with smaller effect sizes ($\bar{d} = -0.49$) exhibited weaker predictive power (mean $\bar{r} = 0.10$). The Gordon parcellation scheme yield very similar patterns (Fig. S5).

3.5. Load-effect sizes indicates parcels' multivariate predictive power

The prior set of results suggests: (a) that the magnitude of WM load-related effects provide clear information regarding the power of parcels to predict between-subjects WM performance effects; and (b) that this effect might be stronger for load-activated than load-deactivated parcels. The next set of analyses examined this question more directly, while also switching to a multivariate approach. Specifically, multivariate approaches enable greater dimensionality reduction, while also testing whether pooling data from multiple parcels achieves a concomitant potential increase in predictive power. We first binned load-sensitive parcels according to their load-related effect sizes, and then estimated the relative predictive power of each parcel bin using an iterative machine learning framework and permutation-based statistical inference (see Method: Building multivariate predictive models of WM load-related effects). The results show that for load-activated parcels, a larger load-related effect size was strongly associated with better predictive power for both in-scanner and out-of-scanner task performance, which was statistically confirmed using linear trend analyses (Fig. 5, Fig. S6). Specifically, parcel bins were rank ordered in terms of linearly increasing effect size (0–6 for load-activated bins) and were then used as independent variables to explain the observed mean of predictive accuracies (in-scanner measures: $b = 0.047$, permutation test: $p < 0.001$; out-of-scanner measures: $b = 0.025$, permutation test: $p < 0.001$). Furthermore, at the largest effect size bins (> 1.1), the predictive power tended to be greater than that observed in univariate analyses (i.e., above 0.4 for in-scanner and 0.2 for out-of-scanner predictions). A qualitatively distinct pattern was observed for load-deactivated parcels. There was no significant linear relationship observed for either in-scanner ($b = 0.02$, permutation test: $p = 0.064$) or out-of-scanner behavioral measures ($b = 0.003$; permutation test: $p = 0.33$). Moreover, we found that for a matched level of load-related effect size, load-activated parcels tended to have significantly higher power than load-deactivated parcels in predicting both in- and out-of-scanner behavioral measures. Specifically, when predicting both the in-scanner 2-back task performance and the out-of-scanner list-sorting task performance, parcels with matched load-related effect sizes between 0.5 and 0.9 (the maximum effect size for load-deactivated parcels), significantly greater predictive power was obtained in the load-activated bins relative to the load-deactivated bins (permutation test: all $p < 0.05$).

3.6. Using load-effect size to guide feature selection

The preceding analyses point to the utility of load-related effect size, multivariate analyses, and load-activated parcel sets when predicting between-subjects variation in WM performance. As a final analysis, we directly tested the hypothesis that load-related effect size can serve as an effective guide to select the most useful features (parcels) when

building predictive models. In particular, we tested models built in an iterative manner, in which features were added to the model successively, according to their ranked load-related effect size (i.e., starting from the parcel with the largest effect size and continuing in descending order). To examine the predictive gain of using load-related effect size as a feature selection guide, for each model, we compared performance to a null model, in which the same number of features (parcels) were selected at random, using a permutation-based approach for statistical inference (see Method: Using load-related effect size for feature selection). Fig. 6a shows that when predicting in-scanner working memory performance, the model accuracy plateaued at around $r = 0.53$ (R^2 = 28% variance explained) when the top 30 parcels ranked by load-effect size were included as predictive features. These top 30 parcels were almost exclusively in fronto-parietal regions contained within the control and dorsal-attention networks (24 out of 30), re-emphasizing the importance of these regions and networks in both within- and between- subjects WM effects (Fig. 6c). Importantly, when using load-related effect size to guide feature selection, the model performance was significantly higher than the same number of parcels were selected at random. In particular, the performance of the load-related models consistently surpassed the 95% cutoff of the null distribution at most feature levels. Both in-scanner as well as out-of-scanner predictive models were tested using both parcellation schemes in order to assess the generalizability of the approach (Fig. S7).

Interestingly, however, when predicting out-of-scanner working memory performance, the model accuracy reached its peak at around when only the top 5 parcels, ranked by load-related effect size, were included as predictive features, with an asymptotic value of $r = 0.27$ (R^2 = 7.3% variance explained; Fig. 6a). It is worth noting that although this value is lower than the in-scanner models, it is also consistent in demonstrating the advantage of using load-related effect size. This is shown by the 95% cutoff of the null distribution never surpassing the observed model performance with up to 5 features. More importantly, the asymptotic level of model performance when predicting both in- and out-of-scanner performance was consistently achieved when utilizing only the top load-related effect size parcels as predictive features. Indeed, when all available predictive features were used, the predictive accuracy was lower, particularly for out-of-scanner prediction (Fig. 6b), which confirms the importance of the utility of selective features, and conversely, the potential vulnerability to over-fitting when using more expansive models. Lastly, this analysis also confirms the value of multivariate relative to univariate models as the multivariate predictive power was greater than even the top univariate predictive parcel (Fig. 6b).

4. Discussion

The goal of the current study was to test the utility of whole-brain parcellation as a dimension reduction approach from which to systematically investigate the relationship of within-subject (load-related) to between-subjects (individual differences) WM variation. This examination yielded a number of important findings. First, parcel-based analyses appear to be an effective form of dimension reduction, in that they yielded high WM load-related effect sizes while also retaining clear specificity to well-established brain regions and networks (frontoparietal control and dorsal attention). Second, a tight coupling was observed between the strength of within- and between-subject effects, particularly for regions showing

a load-activated pattern (high load > low load activation), with large load-related effect sizes predicting stronger brain-behavior relationships. Third, the strength of neural load effects provided an excellent guide to the power of parcels in predicting between-subjects WM performance variation, though this pattern was much stronger when comparing load-activated with load-deactivated (low load > high load) parcels. Fourth, we validated that this property of load-activated parcels can be effectively leveraged as a heuristic guide to feature selection, to build more powerful predictive models of both in-scanner and out-of-scanner WM performance. We elaborate on each of these findings and their implications below.

4.1. Toward using predefined parcellation scheme for dimension reduction

Previous analyses have highlighted the problems of having small sample sizes for neuroimaging research (Button et al., 2013; Desmond and Glover, 2002; Nee, 2019; Szucs and Ioannidis, 2020; Turner et al., 2019) For example, Poldrack et al. (2017) pointed out that at the median sample size of fMRI studies (at around the year 2016), a relatively large effect size of Cohen's $d = 0.75$ is required to have adequate statistical detection power (i.e., > 80%), but yet that the typical voxel/vertex-level effect sizes for the N-back WM task, as measured with the HCP data, was under $d = 0.5$. In the current study, we first replicated the finding reported in Poldrack et al. (2017), and found that the mean vertex-level effect size for the most engaged network (i.e., FPN control network) was barely above $d = 0.5$, with only ~30% vertices surpassing the cutoff of $d = 0.75$. Next, we performed the same effect size analyses at the level of parcels rather than voxels/vertices. This change yielded a qualitative difference from the pattern of results reported above, and with that reported in Poldrack et al. (2017). Specifically, when conducting the same analyses at the parcel level, we found that over half of the parcels within the frontoparietal control network showed large effect sizes (and ~70% of the parcels in this network had effect sizes greater than 0.75). This finding suggests that even with the typical sample sizes used in standard small-scale, single-lab fMRI studies of WM (N ~ 30), parcel-based analyses should be adequately powered to reliably identify load effects. Conversely, we found that in brain networks not typically associated with WM load effects (Visual, Limbic), none of the parcels achieved these large effect sizes. This pattern indicates that parcel-based analyses are highly effective in strongly increasing sensitivity while retaining specificity. Thus, the current results point to the utility of parcel-based analyses as a powerful dimension reduction approach for neuroimaging studies.

Parcel-based analyses also fare well when compared to other dimension reduction approaches that have been frequently adopted in the neuroimaging literature. In particular, component analyses (e.g., PCA, ICA) are a preferred data-driven approach to dimension reduction, as they can be used to extract the latent sources from all brain voxels (or vertices) simultaneously, reducing the feature counts from a hundred thousand to typically less than a hundred (and often on the order of a dozen or two), while still capturing most of the variance in the original dataset. The loadings from these extracted components are then used to examine the relationship between neural activity and behavioral performance, mapping the anatomical location of each latent source (Egli et al., 2018; Sripada et al., 2020). Although component-based approaches are powerful for reducing dimensionality and building predictive models, they also have significant limitations. In particular, component

maps require significant effort to generate, and are difficult to report and communicate in a compact and replicable manner. In contrast, with predefined parcellation schemes, identified neural substrates from one study can be easily communicated (by parcel ID) to other researchers, and as such can easily be utilized by these researchers in follow-up analyses or new studies (e.g., Tables 3, S2). Moreover, along with previous literature, it has been suggested that analyses with predefined parcellation schemes output meaningful results, not only for predictive modeling, but also for applications involving univariate analysis, multivariate analyses (Etzel et al., 2020) or connectivity analyses (Cole et al., 2016, 2021). Indeed, the approach we utilized here could easily be adopted with other domains examined in the HCP, such as language, emotion, and reward processing, to verify the expected gain in effect size and utility for task-related analyses of interest. As such, we strongly advocate for the wider adoption of predefined parcellation schemes as the dimension reduction approach of choice for task-related neuroimaging data analyses.

Nevertheless, it is important to note that the use of predefined parcellation schemes in task fMRI analyses is still nascent, and as such there are many potential complexities that have not been adequately investigated. First, it is important to acknowledge that individual differences in whole brain network architectures have been observed (Mueller et al., 2013; Seitzman et al., 2019). Thus, a predefined parcellation scheme might be more applicable to group-averaged data analyses with big sample size (like the current study) compared to clinical oriented case studies, which may be benefit more from individual-specific parcellation techniques (e.g., Gordon et al., 2017; Wang et al., 2015). Second, many parcellation schemes are defined from whole-brain resting-state functional connectivity profiles, yet it is still not well understood the degree to which brain network structure might change across resting and task states. For example, the default mode network (DMN) defined by the Schaefer parcellation includes a few highly load-activated parcels from the lateral prefrontal cortex (Table 3), which are usually excluded from the conventional DMN when studying working memory (e.g., Čeko et al. 2015). Recent work has suggested that these parcels might actually function as bridges connecting the DMN with other large scale networks, such as the frontoparietal network (Gordon et al., 2020). Further investigation is required to understand the exact role of these parcels during WM task performance. Third, the publication of new parcellation schemes has greatly proliferated in recent years, with various approaches and constraints incorporated into the parcel generation algorithm. As such, a clear “gold standard” has yet to emerge regarding which parcellation scheme to use, or even of the granularity of parcellation (e.g., 100, 400 or 1000 parcel schemes within Schaefer). In fact, it seems likely that the types of brain data one works with (e.g., resting vs. task state data) and the type of analyses one performs (e.g., network analyses vs. multivariate pattern analyses) may have an important impact on which parcellation scheme is most appropriate or effective. Even the issue of how evaluate various parcellation schemes in terms of benchmarks or metrics is one that is only just now finding its way into the literature (Dadi et al., 2020; Zhi et al., 2021). Therefore, for the current study, we replicated all analyses with two separate parcellations (Gordon, Schafer). The results were strongly consistent, which provides strong reassurance regarding the generality of our conclusions. Nevertheless, to provide even more generality, additional parcellation schemes would need to be tested, albeit with highly diminishing returns. As such, more work is needed to provide

researchers with the tools and information to enable selection of the most appropriate parcellation scheme for a particular neuroimaging analysis.

4.2. Tight coupling between within-subject and between-subject WM effects

In the cognitive neuroscience of WM, an important yet unanswered question is: to what degree do the brain regions that support WM (i.e., those showing large neural load effects) overlap with brain regions that are most sensitive to individual differences in WM (i.e., regions showing strong neural-behavioral correlations; Yarkoni and Braver, 2010)? One possibility is that the neural mechanisms used for WM maintenance and manipulation are distinct from those that translate these processes into successful behavioral actions (Egli et al., 2018; Gray et al., 2003; Osaka et al., 2003; Postle et al., 2001). Kane (2003) used a car metaphor to illustrate the potentially dissociable relationship between WM load effects and individual variation in WM performance. In this metaphor, most cars use a similar basic braking mechanism (involving pads, drums, master cylinders, etc.) and the components of this mechanism can be identified by analyzing cars stopping (vs. accelerating; within-subject WM load effects), but variation in stopping distance (between-subjects WM variation) might arise from wholly different mechanisms (aerodynamics, tire balance, weight, pedal placement, etc.), which is only revealed by analyzing different cars that vary on these mechanisms. Yet an alternative, and highly plausible, possibility is that variation in the functioning or efficacy of the brain network utilized for WM maintenance and manipulation is also the primary source of individual variation in WM task performance (Gray et al., 2003; Lee et al., 2006). The primary goal of the current study was to directly compare brain regions identified by these two experimental approaches, in a systematic way, by leveraging the large sample-size of the HCP data set to conduct analyses in a rigorous and well-powered manner.

Our results clearly favor the overlapping neural mechanisms account of WM function. In particular, we observed three key findings that support this interpretation. First, the brain regions showing significant within- and between-subjects WM effects showed a close anatomic overlap, targeting frontoparietal and dorsal attention networks. Second, when directly comparing the magnitude of both types of effects, a tight coupling was observed, in that parcels exhibiting the highest within-subject effect sizes also tended to show the strongest correlations with behavioral performance (i.e., high between-subject effect sizes). Finally, there was no easily observable spatial gradient or pattern in the parcels that exhibited stronger within- or between-subject effects. This finding counters the idea that there is clear dissociation in WM neural mechanisms with prefrontal showing stronger load-related effects and parietal showing stronger sensitivity to individual variation (c.f., Egli et al. 2018; Yarkoni and Braver 2010). Nevertheless, we did observe subtle biases in some brain networks related to greater sensitivity to one type of variation. Specifically, as a load-activated network with large load-related effect sizes, parcels in the dorsal-attention network tended to show greater sensitivity to individual differences relative to the magnitude of load-related effect size. Conversely, in the somatomotor network, which mostly contained parcels exhibiting a load-deactivated pattern, there was relative insensitivity to individual differences. These results are consistent with our overall finding that load-deactivated

regions were less predictive of individual differences in WM task performance relative to load-activated regions, as discussed further below.

It is important to consider potential reasons why our results might appear discrepant from prior work emphasizing dissociations in within-vs. between-subjects WM effects. One factor relates to sample size. Specifically, having sufficient power to detect between-subjects WM effects requires a much larger sample than is necessary for detecting within-subject effects (Braver et al., 2010; Yarkoni and Braver, 2010). As discussed above, it is only more recently that fMRI studies have provided sample sizes necessary for conducting a comprehensive analysis of between-subjects WM variation. Thus, most of the prior published work on this topic either had exceedingly low power for replication (Bunge, 2001; Todd and Marois, 2005), or instead focused on isolated *a priori* regions of interest (e.g., Osaka et al. 2003). By taking advantage of the large sample size of the HCP and a whole-brain parcellation scheme, the current study provides a powerful approach by which to resolve prior discrepancies in the literature. It is also the case that the discrepancy might be more apparent than real. Indeed, in a supplementary analysis, we used the intraclass correlation coefficient (ICC) as an independent metric of between-subject variation. As expected, the ICC was highly negatively correlated with load-effect size, strongly confirming the tradeoff between these two types of effects (Fig. S9a). Nevertheless, ICC proved to be a much weaker predictive indicator of brain-behavior relationships than load-effect size (Fig. S9b), supporting the general conclusions of our primary analyses.

Nevertheless, the current findings are subject to two important limitations. First, due to the nature of the N-back task design in HCP, we were only able to measure the load-effect and individual differences for 2-back performance (relative to 0-back). Working memory tasks with higher working memory load demands would provide the ability to detect load-related effects and their relationship to WM task performance in a richer and more nuanced manner. For example, in recent work, Lamichhane et al. (2020) used an N-back design that involved 6 parametric levels of WM load. The analyses utilized several different metrics to describe load-related effects and relate these to individual differences in WM performance. These analyses identified a single region in left dorsolateral PFC, for which load-related effects predicted WM performance; though again the sample size in that study was too small (N ~50) to conduct a comprehensive whole-brain analysis of between-subjects WM effects. An ideal design would be one that examined the relationship within- and between-subjects WM variation using a parametric WM load manipulation, rather than with a single high-load level. Such experimental designs can be achieved using N-back tasks with multiple parametric levels (e.g., Lamichhane et al., 2020) or other WM tasks that involve a wide range of load manipulations such as the Self Ordered memory task (Van Snellenberg et al., 2015) and the Sternberg Item Recognition task (Rypma et al., 2002). Yet a second limitation to consider is that even with the sample-size of the HCP (N ~ 1000), it is possible that this dataset is still too small to generate precise and generalizable estimates of predictive power. In particular, a recent study has suggested that the brain-wide associations we observed here (i.e., the correlation between neural activity and WM performance) may still be subject to strong effect size inflation and mis-estimation due to sampling variability, that does not resolve until samples are at least twice as large as the HCP dataset (Marek et al., 2020). However, it is worth noting that Marek et al. (2020) examined the brain-wide

associations using resting state functional connectivity as the neural measure to predict individual differences in behavioral task performance. Some studies have suggested that neural measures derived from cognitively demanding task states will have more robust and consistent predictive power to capture individual differences in behavioral performance (Sripada et al., 2020).

4.3. Magnitude and direction of load-related effect size indicates predictive power

A correlational approach (or simple linear regression) measures the degree to which the neural activity of a parcel can be used to *explain* individual differences in behavioral performance. However, correlational approaches do not necessarily indicate whether such a parcel can be reliably used to *predict* out-of-sample data (Yarkoni and Westfall, 2017). Consequently, in addition to standard correlational analyses, we used cross-validation approaches to directly explore how predictive power for between-subjects WM effects varied as a function of load-related effect size. To explore this issue more thoroughly, both univariate and multivariate approaches were used to measure predictive power. Similar to what has now been observed in many fMRI analyses, we found that multivariate models substantially outperformed univariate ones in predicting both in- and out-of-scanner tasks (Fig. 6b, Fig. S7). Importantly, however, both approaches converged in suggesting that the predictive power of a parcel (or a set of parcels) increases as the load-related effect size increases.

An additional key finding in these analyses was that the univariate and multivariate models both provided evidence load-activated parcels had reliably greater predictive power than load-deactivated parcels (Fig. 5). Note that this difference in predictive power between load-activated and load-deactivated parcels was not an artifact of controlling (the absolute value of) load-related effect size or the number of predictive features (parcels). In fact, supplemental analyses revealed that the general pattern still held even when comparing the predictive power of all load-deactivated regions to load-activated regions (Fig. S8). We speculate that there are two possible reasons why load-deactivated parcels might exhibit less predictive power than load-activated. First, regions that typically exhibit a load deactivation pattern (e.g., the default mode network) appear to be sensitive to more general and non-specific factors, such as mind-wondering, arousal, and fatigue, that may only peripherally or indirectly contribute to task performance. Second, while load-activated regions included parcels mostly from brain networks that are well-established to support higher cognitive functions, such as WM and attention control (e.g., Control/Frontoparietal and Dorsal Attention networks), load-deactivated parcels were primarily associated with perceptually-oriented regions, such as those that make up the Visual and Somatomotor networks. Together, the anatomic location and functional properties of load-deactivated parcels are consistent with the finding that these parcels were less predictive of task performances relative to load-activated ones.

However, the current findings do diverge from Satterthwaite et al. (2013), who reported results from a large sample N-back study indicating that load-activated and load-deactivated regions have similar predictive power. Two key differences between the current study and Satterthwaite et al. (2013) may account for this discrepancy. First, Satterthwaite et al. (2013)

defined load-deactivated regions of interest (ROI) only from the default mode network. In contrast, the current study found that load-deactivated regions could also be located in perceptually-relevant brain networks (Visual / Somatomotor). Second, Satterthwaite et al. (2013) defined ROIs based on the weights of another cross-validated predictive model, leading to ROIs of high predictive power. In the current study, parcels were selected according to load effect size, rather than predictive power. Additionally, we want to stress that load-deactivated parcels were defined according to a contrast between two load conditions (2-back, 0-back), rather than between 2-back and a fixation baseline. We believe this contrast was appropriate, given our primary interest in identifying brain regions whose activity was specifically modulated by WM load. However, based on this contrast, we did not differentiate regions showing distinct deactivation profiles (i.e., baseline > 0-back > 2-back vs. 0-back > 2-back > baseline). Although beyond the scope of the current paper, identifying whether such dissociable patterns of deactivation are present, and whether they exhibit distinct profiles of predictive power for cognitive task performance, would be an interesting question for future work.

4.4. Load effect being a good indicator for feature selection

A primary goal of cognitive neuroscience research is to find effective neural markers that can reliably *predict* individual differences in behavior. A number of recent studies have attempted to build predictive models from whole-brain task-related activity patterns in order to predict WM performance (Egli et al., 2018; Pornpattananangkul et al., 2020) and general cognitive ability (Sripada et al., 2020). However, despite their success in predicting behavior, our results argue that including neural activity from the whole brain might be suboptimal. Instead, selectively choosing predictive features for the model may further maximize its predictive accuracy (Fig. 6b). Specifically, when including all 400 features (Schaefer parcellation scheme), the model predictive accuracy for the in-scanner (N-back) WM task was around $r = 0.49$, which is around the same accuracy level compared to previous models (Pornpattananangkul et al., 2020; Satterthwaite et al., 2013). However, our results show that the model achieved a better performance when only the top 30 features ranked based on load-effect size were used ($r = 0.53$). The same pattern was observed when predicting out-of-scanner behavioral measures and using the other parcellation scheme (Figure S7). Together, these results suggest that predictive models for WM performance may benefit from feature selection approaches, particularly when these are guided by a functionally relevant principle.

More specifically, the current study results provide a clear indication that load-effect size can serve as a principled basis from which to guide feature (i.e., parcel) selection. We conducted permutation-based analyses which compared feature selection based on WM load relative to random selection of an equivalent number of features. These analyses revealed that WM load-based feature selection can almost always lead to the best model performance, given any predictive feature size (Fig. 6a). Indeed, in supplementary analyses, we found that the load-based approach to feature selection fared very well when compared to a standard machine-learning approach to feature selection (the forward stepwise algorithm; Fig. S10) that is designed to maximize predictive power, but in a ‘blind’ data-driven fashion (i.e., rather than according to interpretable principles). Importantly, only the top 30 features

ranked by load-effect size were needed for the model to achieve optimal performance. When predicting out-of-scanner working memory performance, it seems that a load-based feature selection principle had little contribution when more than 10 features were selected. However, the results did show that with load-effect size as an indicator, a model could achieve peak performance (better than when all 400 features were included) when only the top 5 load-related features were selected. Together, the current study provides evidence that feature (parcel) selection is an important step toward building the optimal predictive model for individual differences in working memory function, and that load-effect size is a good indicator to guide the feature selection process in a principled manner.

4.5. Constraints on generality

As described above, there are a few limitations of the current study which provide important constraints regarding the conclusions that can be drawn from it. In addition to the fact that the study involved only a categorical (high / low) rather than parametric manipulation of WM load, and may still have had an insufficient sample size to yield robust results, it is important to consider the HCP sample, the use of the N-back task, the focus on activation rather than connectivity effects, and the stimuli themselves as providing constraints on generality. We address each of these points briefly. The HCP sample included only healthy young adults (22–35), and thus the results may not generalize well to developmental, aging, or other populations suffering from clinical impairment. The N-back task is only one out of many potential experimental paradigms used to probe WM function (Wilhelm et al., 2013). Further, although it is one of the most popularly used in neuroimaging studies, the N-back task suffers from many well-known limitations as a pure probe of both WM maintenance functions and also individual differences (Kane et al., 2007). Even within the N-back, the HCP task is relatively non-representative, as prior work has more frequently tended to focus on verbal (letters, words) or spatial materials, rather than the mostly non-verbal stimuli (e.g., places, faces) used in the HCP, with some clear anatomic distinctions observed across these stimulus factors (Owen et al., 2005). Thus, it is quite plausible that the pattern observed here may not generalize to other N-back variants. Finally, although some recent studies have focused on the potential for task-based activation patterns to be used as robust predictors of individual differences (Etzel et al., 2020; Satterthwaite et al., 2013; Sripada et al., 2020) more attention has been given towards connectivity approaches, both resting-state and task-based, in this domain. As mentioned above, it is not yet clear whether the critical factors that impact detection of connectivity-based individual difference analyses (Finn et al., 2015; Marek et al., 2020), generalize to analysis approaches, such as this one, that involve activation-based metrics. We add this section to remind researchers that all these factors need to be kept in mind when drawing implications from the current work towards future studies (Simons et al., 2017; Yarkoni, 2020).

5. Conclusion

The current study jointly informs two distinct issues within cognitive neuroscience: 1) the degree of overlap in the neural substrates of WM function related to within-subjects (e.g., load-related) and between-subjects (individual differences) effects; and 2) the neural indices that show the greatest predictive power for detecting individual differences in cognitive

function. Our findings demonstrate that, at least within the HCP dataset and N-back task variants used in that study, neural load effects are tightly linked with individual variation in cognitive task performance, and as such, can be used as the basis for feature selection to build predictive models of individual differences in WM function (as well as other indicators of general cognitive ability). In so doing, the current work highlights the utility of large sample datasets, whole-brain parcel-based approaches, and the use of informed feature selection in neurally-based predictive modeling. As such, our findings provide a strong foundation for future studies that can expand upon these efforts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding and Acknowledgement

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 Mc-Donnell Center for Systems Neuroscience at Washington University. Additional support provided by NIH R37 MH066078 to TSB. The authors thank Joset A. Etzel for helpful discussions on the analyses and feedback on the manuscript.

References

Baddeley A, 2012. Working memory: theories, models, and controversies. Annu. Rev. Psychol 63, 1–29. doi: 10.1146/annurev-psych-120710-100422. [PubMed: 21961947]

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, Nolan D, Bryant E, Hartley T, Footer O, Bjork JM, Poldrack R, Smith S, Johansen-Berg H, Snyder AZ, Van Essen DC, 2013. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80, 169–189. doi: 10.1016/j.neuroimage.2013.05.033. [PubMed: 23684877]

Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC, 1997. A parametric study of prefrontal cortex involvement in human working memory. Neuroimage 5 (1), 49–62. doi: 10.1006/nimg.1996.0247. [PubMed: 9038284]

Braver TS, Cole MW, Yarkoni T, 2010. Vive les differences! Individual variation in neural mechanisms of executive control. Curr. Opin. Neurobiol 20 (2), 242–250. doi: 10.1016/j.conb.2010.03.002. [PubMed: 20381337]

Buckner RL, Andrews-Hanna JR, Schacter DL, 2008. The brain's default network: anatomy, function, and relevance to disease. Ann. N. Y. Acad. Sci 1124, 1–38. doi: 10.1196/annals.1440.011. [PubMed: 18400922]

Bunge SA, 2001. Prefrontal regions involved in keeping information in and out of mind. Brain 124 (10), 2074–2086. doi: 10.1093/brain/124.10.2074. [PubMed: 11571223]

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR, 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosc 14 (5), 365–376. doi: 10.1038/nrn3475.

Cairo TA, Liddle PF, Woodward TS, Ngan ETC, 2004. The influence of working memory load on phase specific patterns of cortical activity. Cognit. Brain Res 21 (3), 377–387. doi: 10.1016/j.cogbrainres.2004.06.014.

Čeko M, Gracely JL, Fitzcharles MA, Seminowicz DA, Schweinhardt P, Bush-nell MC, 2015. Is a responsive default mode network required for successful working memory task performance? J. Neurosc 35 (33), 11595–11605. doi: 10.1523/JNEUROSCI.0264-15.2015.

Cole MW, Ito T, Bassett DS, Schultz DH, 2016. Activity flow over resting-state networks shapes cognitive task activations. Nat. Neurosci 19 (12), 1718–1726. doi: 10.1038/nn.4406. [PubMed: 27723746]

Cole MW, Ito T, Cocuzza C, Sanchez-Romero R, 2021. The functional relevance of task-state functional connectivity. J. Neurosc doi: 10.1523/JNEUROSCI.1713-20.2021, JN-RM-1713-20.

Cooper SR, Jackson JJ, Barch DM, Braver TS, 2019. Neuroimaging of individual differences: a latent variable modeling perspective. Neurosc. Biobehav. Rev 98, 29–46. doi: 10.1016/j.neubiorev.2018.12.022.

Dadi K, Varoquaux G, Machlouzarides-Shalit A, Gorgolewski KJ, Wassermann D, Thirion B, Mensch A, 2020. Fine-grain atlases of functional modes for fMRI analysis. Neuroimage 221, 117126. doi: 10.1016/j.neuroimage.2020.117126. [PubMed: 32673748]

Desmond JE, Glover GH, 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J. Neurosci. Methods 118 (2), 115–128. doi: 10.1016/s0165-0270(02)00121-8. [PubMed: 12204303]

Egli T, Coynel D, Spalek K, Fastenrath M, Freytag V, Heck A, Loos E, Auschra B, Papassotiropoulos A, de Quervain DJF, Milnik A, 2018. Identification of two distinct working memory related brain networks in healthy young adults. eNeuro doi: 10.1523/ENEURO.0222-17.2018, ENEURO.0222–17.2018

Engle RW, Kane MJ, Tuholski SW, 1999. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In: Models of Working memory: Mechanisms of Active Maintenance and Executive Control Cambridge University Press, pp. 102–134. doi: 10.1017/CBO9781139174909.007.

Etzel JA, Braver TS, 2013. MVPA permutation schemes: permutation testing in the land of cross-validation. In: Proceedings of the International Workshop on Pattern Recognition in Neuroimaging, pp. 140–143. doi: 10.1109/PRNI.2013.44.

Etzel JA, Courtney Y, Carey CE, Gehred MZ, Agrawal A, Braver TS, 2020. Pattern similarity analyses of frontoparietal task coding: individual variation and genetic influences. Cereb. Cortex 30 (5), 3167–3183. doi: 10.1093/cercor/bhz301. [PubMed: 32086524]

Feredoes E, Postle BR, 2007. Localization of load sensitivity of working memory storage: quantitatively and qualitatively discrepant results yielded by single-subject and group-averaged approaches to fMRI group analysis. Neuroimage 35 (2), 881–903. doi: 10.1016/j.neuroimage.2006.12.029. [PubMed: 17296315]

Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT, 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci 18 (11), 1664–1671. doi: 10.1038/nn.4135. [PubMed: 26457551]

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC, Jenkinson M WU-Minn HCP Consortium, 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127. [PubMed: 23668970]

Gordon EM, Laumann TO, Adeyemo B, Gilmore AW, Nelson SM, Dosenbach NUF, Petersen SE, 2017. Individual-specific features of brain systems identified with resting state functional correlations. Neuroimage 146, 918–939. doi: 10.1016/j.neuroimage.2016.08.032. [PubMed: 27640749]

Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE, 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cerebral Cortex 26 (1), 288–303. doi: 10.1093/cercor/bhu239, (New York, N.Y.: 1991). [PubMed: 25316338]

Gordon EM, Laumann TO, Marek S, Raut RV, Gratton C, Newbold DJ, Greene DJ, Coalson RS, Snyder AZ, Schlaggar BL, Petersen SE, Dosenbach NUF, Nelson SM, 2020. Default-mode network streams for coupling to language and control systems. In: Proceedings of the National Academy of Sciences, 117, pp. 17308–17319. doi: 10.1073/pnas.2005238117.

Gray JR, Chabris CF, Braver TS, 2003. Neural mechanisms of general fluid intelligence. Nat. Neurosci 6 (3), 316–322. doi: 10.1038/nn1014. [PubMed: 12592404]

Jha AP, McCarthy G, 2000. The influence of memory load upon delay-interval activity in a working-memory task: an event-related functional MRI study. J. Cogn. Neurosci 12 (Supplement 2), 90–105. doi: 10.1162/089892900564091. [PubMed: 11506650]

Just MA, Carpenter PA, 1992. A capacity theory of comprehension: individual differences in working memory. Psychol. Rev 99 (1), 122–149. doi: 10.1037/0033-295X.99.1.122. [PubMed: 1546114]

Kampa K, Mehta S, Chou CA, Chaovalitwongse WA, Grabowski TJ, 2014. Sparse optimization in feature selection: application in neuroimaging. J. Glob. Optim 59 (2), 439–457. doi: 10.1007/s10898-013-0134-2.

Kane MJ, 2003. The intelligent brain in conflict. Trends Cogn. Sci. (Regul. Ed.) 7 (9), 375–377. doi: 10.1016/S1364-6613(03)00193-1.

Kane MJ, Conway ARA, Miura TK, Colflesh GJH, 2007. Working memory, attention control, and the N-back task: a question of construct validity. J. Exp. Psychol. Learn Mem. Cogn 33 (3), 615–622. doi: 10.1037/0278-7393.33.3.615. [PubMed: 17470009]

Lamichhane B, Westbrook A, Cole MW, Braver TS, 2020. Exploring brain-behavior relationships in the N-back task. Neuroimage 212, 116683. doi: 10.1016/j.neuroimage.2020.116683. [PubMed: 32114149]

Lee KH, Choi YY, Gray JR, Cho SH, Chae J-H, Lee S, Kim K, 2006. Neural correlates of superior intelligence: stronger recruitment of posterior parietal cortex. Neuroimage 29 (2), 578–586. doi: 10.1016/j.neuroimage.2005.07.036. [PubMed: 16122946]

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, Conan G, Uriarte J, … Dosenbach NUF, 2020. Towards reproducible brain-wide association studies. BioRxiv doi: 10.1101/2020.08.21.257758, 2020.08.21.257758.

Mitchell DJ, Cusack R, 2008. Flexible, capacity-limited activity of posterior parietal cortex in perceptual as well as visual short-term memory tasks. Cerebral Cortex 18 (8), 1788–1798. doi: 10.1093/cercor/bhm205, (New York, N.Y. 1991). [PubMed: 18042643]

Motes MA, Rypma B, 2010. Working memory component processes: isolating BOLD signal changes. Neuroimage 49 (2), 1933–1941. doi: 10.1016/j.neuroimage.2009.08.054. [PubMed: 19732840]

Mueller S, Wang D, Fox MD, Yeo BTT, Sepulcre J, Sabuncu MR, Shafee R, Lu J, Liu H, 2013. Individual variability in functional connectivity architecture of the human brain. Neuron 77 (3), 586–595. doi: 10.1016/j.neuron.2012.12.028. [PubMed: 23395382]

Nee DE, 2019. FMRI replicability depends upon sufficient individual-level data. Commun. Biol 2 (1), 1–4. doi: 10.1038/s42003-019-0378-6. [PubMed: 30740537]

Osaka M, Osaka N, Kondo H, Morishita M, Fukuyama H, Aso T, Shibasaki H, 2003. The neural basis of individual differences in working memory capacity: an fMRI study. Neuroimage 18 (3), 789–797. doi: 10.1016/S1053-8119(02)00032-0. [PubMed: 12667855]

Owen AM, McMillan KM, Laird AR, Bullmore E, 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum. Brain Mapp 25 (1), 46–59. doi: 10.1002/hbm.20131. [PubMed: 15846822]

Pessoa L, Gutierrez E, Bandettini P, Ungerleider L, 2002. Neural correlates of visual working memory: FMRI amplitude predicts task performance. Neuron 35 (5), 975–987. doi: 10.1016/s0896-6273(02)00817-6. [PubMed: 12372290]

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline JB, Vul E, Yarkoni T, 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosc 18 (2), 115–126. doi: 10.1038/nrn.2016.167.

Pornpattananangkul N, Bartonicek A, Wang Y, Stringaris A, 2020. An omics-inspired elastic net approach drastically improves out-of-sample prediction and regional inference of task-based fMRI [Preprint]. Neuroscience doi: 10.1101/2020.10.21.348367.

Postle BR, Berger JS, Goldstein JH, Curtis CE, D'Esposito M, 2001. Behavioral and neurophysiological correlates of episodic coding, proactive interference, and list length effects in a running span verbal working memory task. Cogn. Affect. Behav. Neurosci 1 (1), 10–21. doi: 10.3758/CABN.1.1.10. [PubMed: 12467100]

Rottschy C, Langner R, Dogan I, Reetz K, Laird AR, Schulz JB, Fox PT, Eickhoff SB, 2012. Modelling neural correlates of working memory: a coordinate-based meta-analysis. Neuroimage 60 (1), 830–846. doi: 10.1016/j.neuroimage.2011.11.050. [PubMed: 22178808]

Rypma B, Berger JS, D'Esposito M, 2002. The influence of working-memory demand and subject performance on prefrontal cortical activity. J. Cogn. Neurosci 14 (5), 721–731. doi: 10.1162/08989290260138627. [PubMed: 12167257]

Rypma B, Prabhakaran V, Desmond JE, Glover GH, Gabrieli JD, 1999. Load-dependent roles of frontal brain regions in the maintenance of working memory. Neuroimage 9 (2), 216–226. doi: 10.1006/nimg.1998.0404. [PubMed: 9927550]

Satterthwaite TD, Wolf DH, Erus G, Ruparel K, Elliott MA, Gennatas ED, Hopson R, Jackson C, Prabhakaran K, Bilker WB, Calkins ME, Loughead J, Smith A, Roalf DR, Hakonarson H, Verma R, Davatzikos C, Gur RC, Gur RE, 2013. Functional maturation of the executive system during adolescence. J. Neurosci 33 (41), 16249–16261. doi: 10.1523/JNEUROSCI.2345-13.2013. [PubMed: 24107956]

Saults JS, Cowan N, 1998. Developmental and individual differences in short-term memory. In: Advances in psychology, 125. North-Holland, pp. 155–196. doi: 10.1016/S0166-4115(98)80006-X.

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT, 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28 (9), 3095–3114. doi: 10.1093/cercor/bhx179. [PubMed: 28981612]

Seitzman BA, Gratton C, Laumann TO, Gordon EM, Adeyemo B, Dworetsky A, Kraus BT, Gilmore AW, Berg JJ, Ortega M, Nguyen A, Greene DJ, McDermott KB, Nelson SM, Lessov-Schlaggar CN, Schlaggar BL, Dosenbach NUF, Petersen SE, 2019. Trait-like variants in human functional brain networks. In: Proceedings of the National Academy of Sciences, 116, pp. 22851–22861.

Simons DJ, Shoda Y, Lindsay DS, 2017. Constraints on generality (COG): a proposed addition to all empirical papers. Perspect. Psychol. Sci 12 (6), 1123–1128. doi: 10.1177/1745691617708630. [PubMed: 28853993]

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saun-ders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM, 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1), S208–S219. doi: 10.1016/j.neuroimage.2004.07.051. [PubMed: 15501092]

Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K, 2020. Toward a “tread-mill test” for cognition: improved prediction of general cognitive ability from the task activated brain. Hum. Brain Mapp 41 (12), 3186–3197. doi: 10.1002/hbm.25007. [PubMed: 32364670]

Szucs D, Ioannidis JP, 2020. Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. Neuroimage 221, 117164. doi: 10.1016/j.neuroimage.2020.117164. [PubMed: 32679253]

Todd JJ, Marois R, 2005. Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. Cogn. Affect. Behav. Neurosci 5 (2), 144–155. doi: 10.3758/CABN.5.2.144. [PubMed: 16180621]

Tulsky DS, Carlozzi N, Chiaravalloti ND, Beaumont JL, Kisala PA, Mungas D, Conway K, Gershon R, 2014. NIH toolbox cognition battery (NIHTB-CB): the list sorting test to measure working memory. J. Int. Neuropsychol. Soc. JINS 20 (6), 599. doi: 10.1017/S135561771400040X. [PubMed: 24959983]

Turner BO, Santander T, Paul EJ, Barbey AK, Miller MB, 2019. Reply to: FMRI replicability depends upon sufficient individual-level data. Commun. Biol 2 (1), 1–3. doi: 10.1038/s42003-019-0379-5. [PubMed: 30740537]

U ğurbil K, Xu J, Auerbach EJ, Moeller S, Vu A, Duarte-Carvajalino JM, Lenglet C, Wu X, Schmitter S, Van de Moortele PF, Strupp J, Sapiro G, De Martino F, Wang D, Harel N, Garwood M, Chen L, Feinberg DA, Smith SM, … Yacoub E, 2013. Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project. Neuroimage 80, 80–104. doi: 10.1016/j.neuroimage.2013.05.012. [PubMed: 23702417]

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, 2013. The WU-minn human connectome project: an overview. Neuroimage 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041. [PubMed: 23684880]

Van Snellenberg JX, Slifstein M, Read C, Weber J, Thompson JL, Wager TD, Shohamy D, Abi-Dargham A, Smith EE, 2015. Dynamic shifts in brain network activation during supracapacity working memory task performance. Hum. Brain Mapp 36 (4), 1245–1264. doi: 10.1002/hbm.22699. [PubMed: 25422039]

Veltman DJ, Rombouts SARB, Dolan RJ, 2003. Maintenance versus manipulation in verbal working memory revisited: an fMRI study. Neuroimage 18 (2), 247–256. doi: 10.1016/s1053-8119(02)00049-6. [PubMed: 12595179]

Wang D, Buckner RL, Fox MD, Holt DJ, Holmes AJ, Stoecklein S, Langs G, Pan R, Qian T, Li K, Baker JT, Stufflebeam SM, Wang K, Wang X, Hong B, Liu H, 2015. Parcellating cortical functional networks in individuals. Nat. Neurosci 18 (12), 1853–1860. doi: 10.1038/nn.4164. [PubMed: 26551545]

Wilhelm O, Hildebrandt AH, Oberauer K, 2013. What is working memory capacity, and how can we measure it? Front. Psychol 4. doi: 10.3389/fpsyg.2013.00433.

Xu Y, Chun MM, 2006. Dissociable neural mechanisms supporting visual short-term memory for objects. Nature 440 (7080), 91–95. doi: 10.1038/nature04262. [PubMed: 16382240]

Yarkoni T, 2020. The generalizability crisis. Behav. Brain Sci 1–37. doi: 10.1017/S0140525X20001685.

Yarkoni T, Braver TS, 2010. Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological issues. In: Gruszka A, Matthews G, Szymura B (Eds.), Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control Springer, pp. 87–107. doi: 10.1007/978-1-4419-1210-7_6.

Yarkoni T, Westfall J, 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci 12 (6), 1100–1122. doi: 10.1177/1745691617693393. [PubMed: 28841086]

Zhi D, King M, & Diedrichsen J (2021). Evaluating brain parcellations using the distance controlled boundary coefficient. BioRxiv, 2021.05.11.443151. 10.1101/2021.05.11.443151

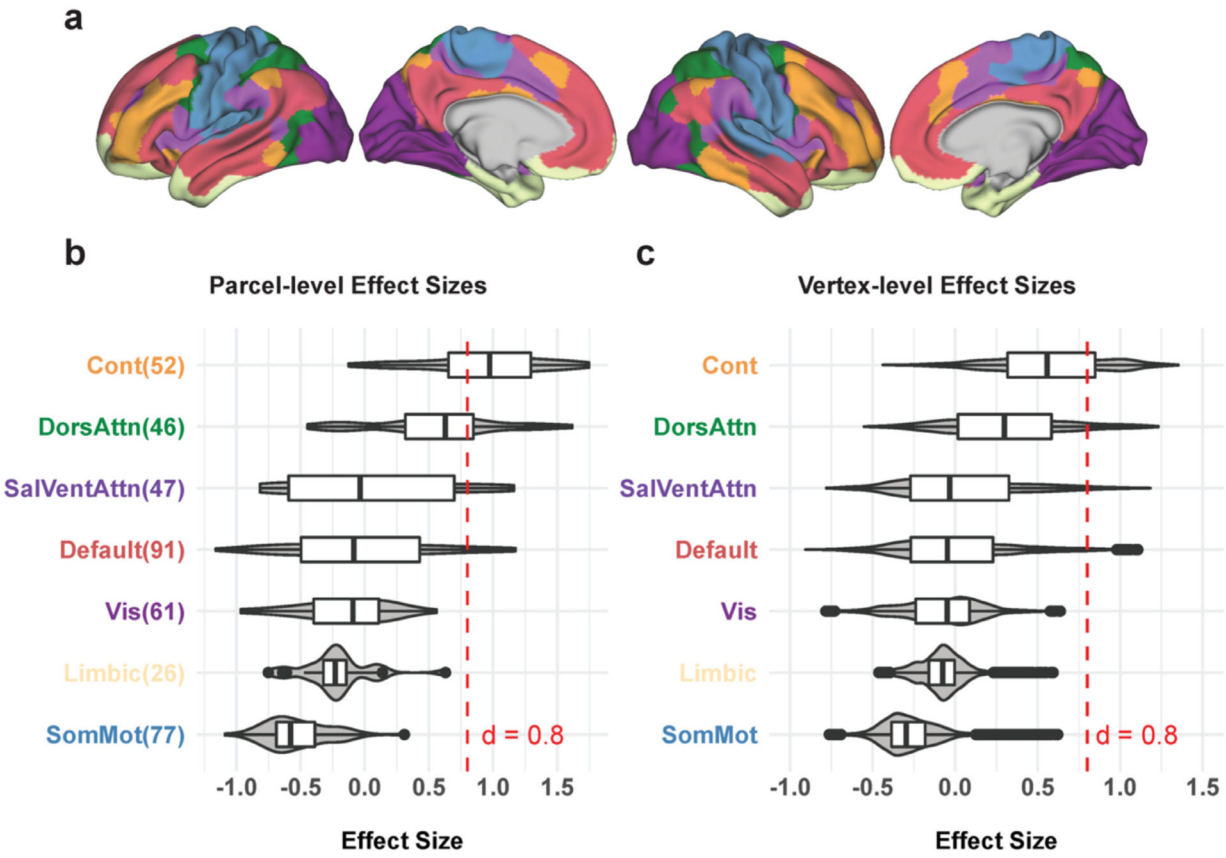

Fig. 1.

Parcel- and vertex-wise load effect sizes with Schaefer parcellation scheme. (**a**) The Schaefer predefined parcellation scheme. The color map shows the respective functional network. (**b,c**) The parcel-level and vertex-level load-related effect sizes, respectively. Vertices and parcels were grouped for each functional network. The number in the parenthesis indicates the number of parcels within each network.

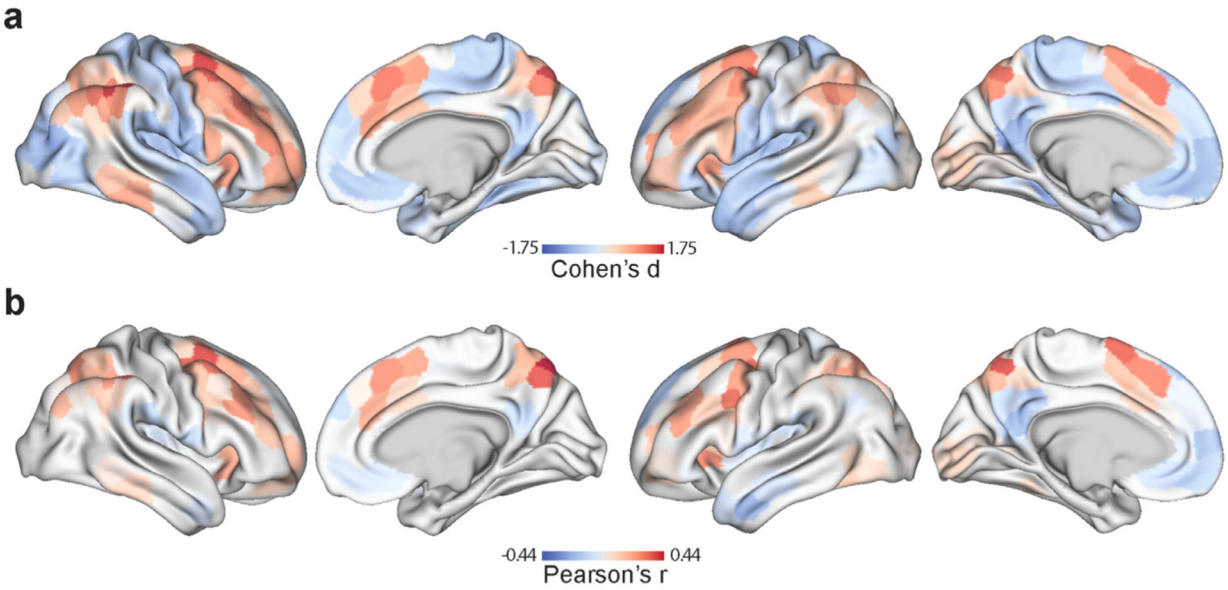

Fig. 2.

Neural correlates of within- and between-subject variations defined using Schaefer parcellation scheme. (**a**) Parcels showing within-subject WM effect (i.e., load-sensitive). The color indicates the load-effect size, with red indicating load-activated parcels and blue load-deactivated parcels. (**b**) Parcels showing between-subjects WM effect (i.e., neural-behavioral correlations). The color indicates the sign and size of the correlation, with red indicating positive and blue negative correlations.

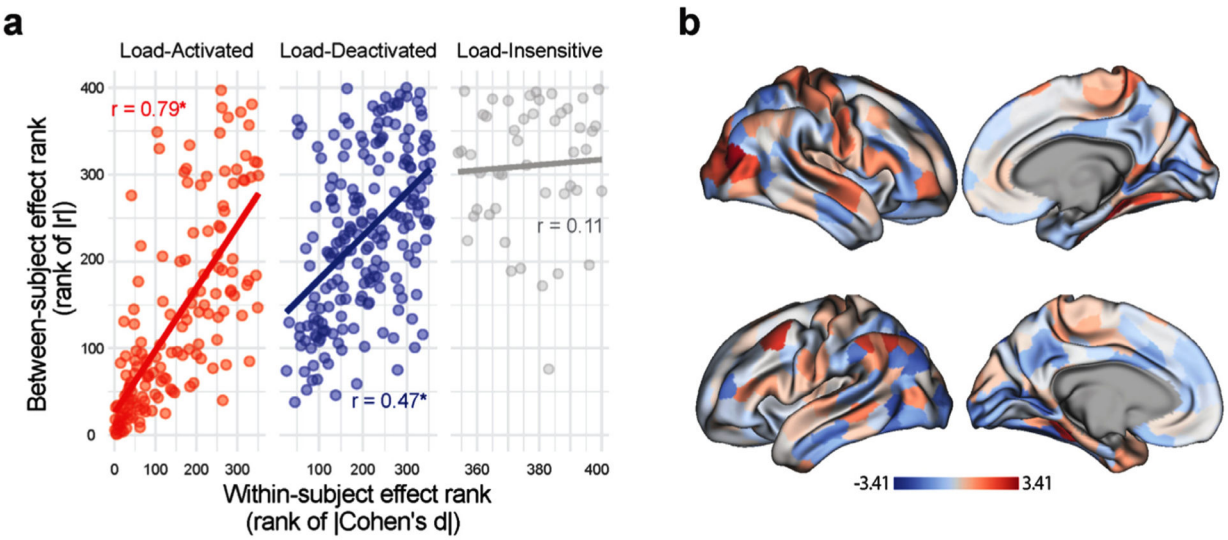

Fig. 3.

(**a**) The relationship between parcel's within-subject effect size and explanatory power, as measured by load-related effect size and brain-behavior correlation using Schaefer parcellation scheme. The labels for x and y axis are the ranked order for within- and between- subject effect sizes, with smaller number indicating higher rank. The correlation between the rank orders for the two effect sizes were computed separately for load-activated, deactivated, and insensitive parcels (* indicates $p < 0.001$). (**b**) The spatial relationship between parcels being more senstive to within-subject effect and those being more sensitive to between-subjects effect. The color map represents the degree to which a parcel is more sensitive to a type of difference. Warm colors indicate a parcel being more sensitive to within-subject variation and cool more sensitive between-subject variation.

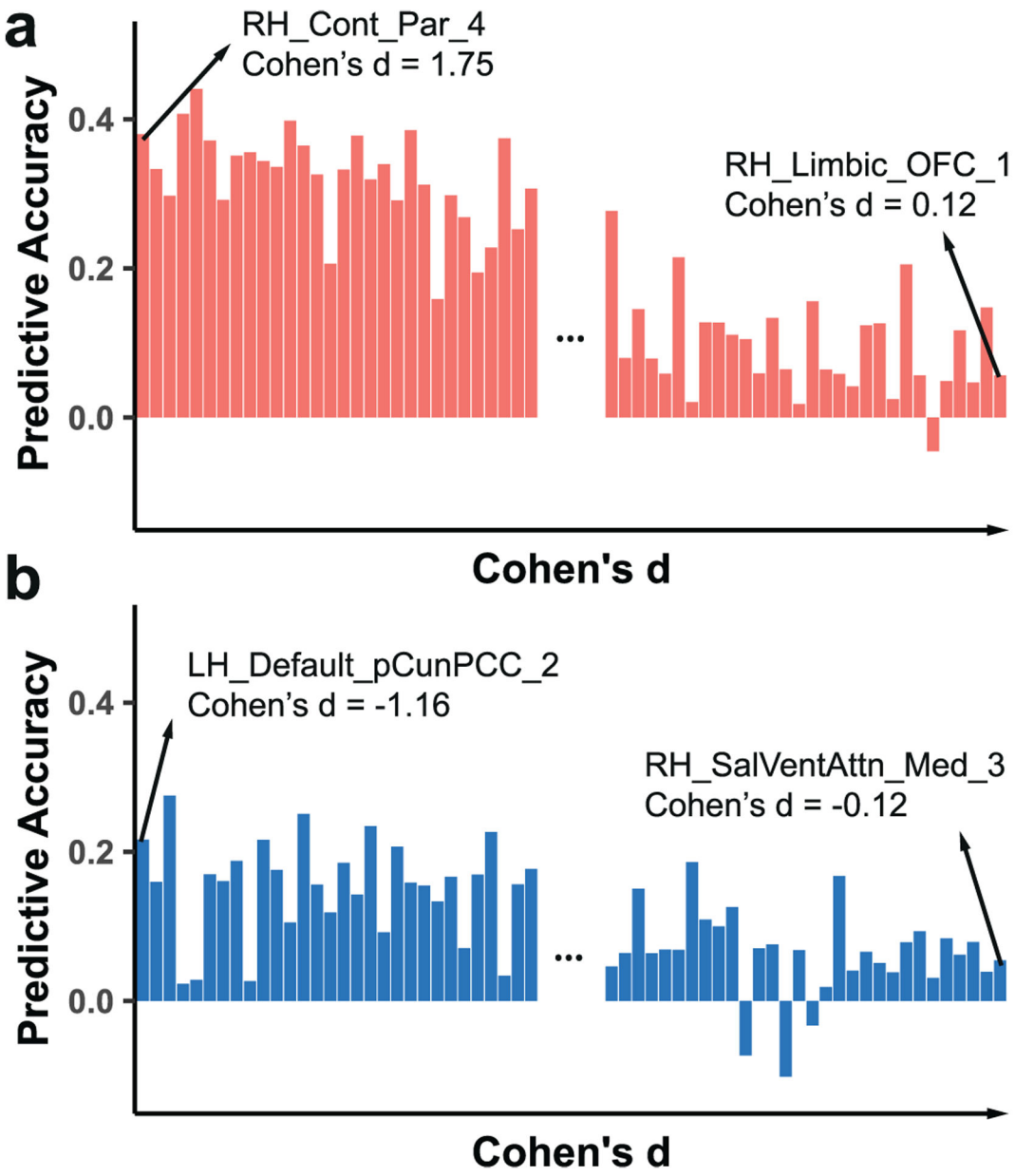

Fig. 4.

The relationship between load-related effect size and predictive accuracy for Schaefer parcels. (**a**) Univariate predictive accuracy of the top and bottom 30 load-activated parcels ranked by load-related effect size (i.e., Cohen's d). (**b**) Univariate predictive accuracy of the top and bottom 30 load-deactivated parcels ranked by the absolute values of the load-related effect size. Predictive power was examined with the left-out samples using cross-validation; predictive accuracy was quantified as the correlation between the predicted and observed scores.

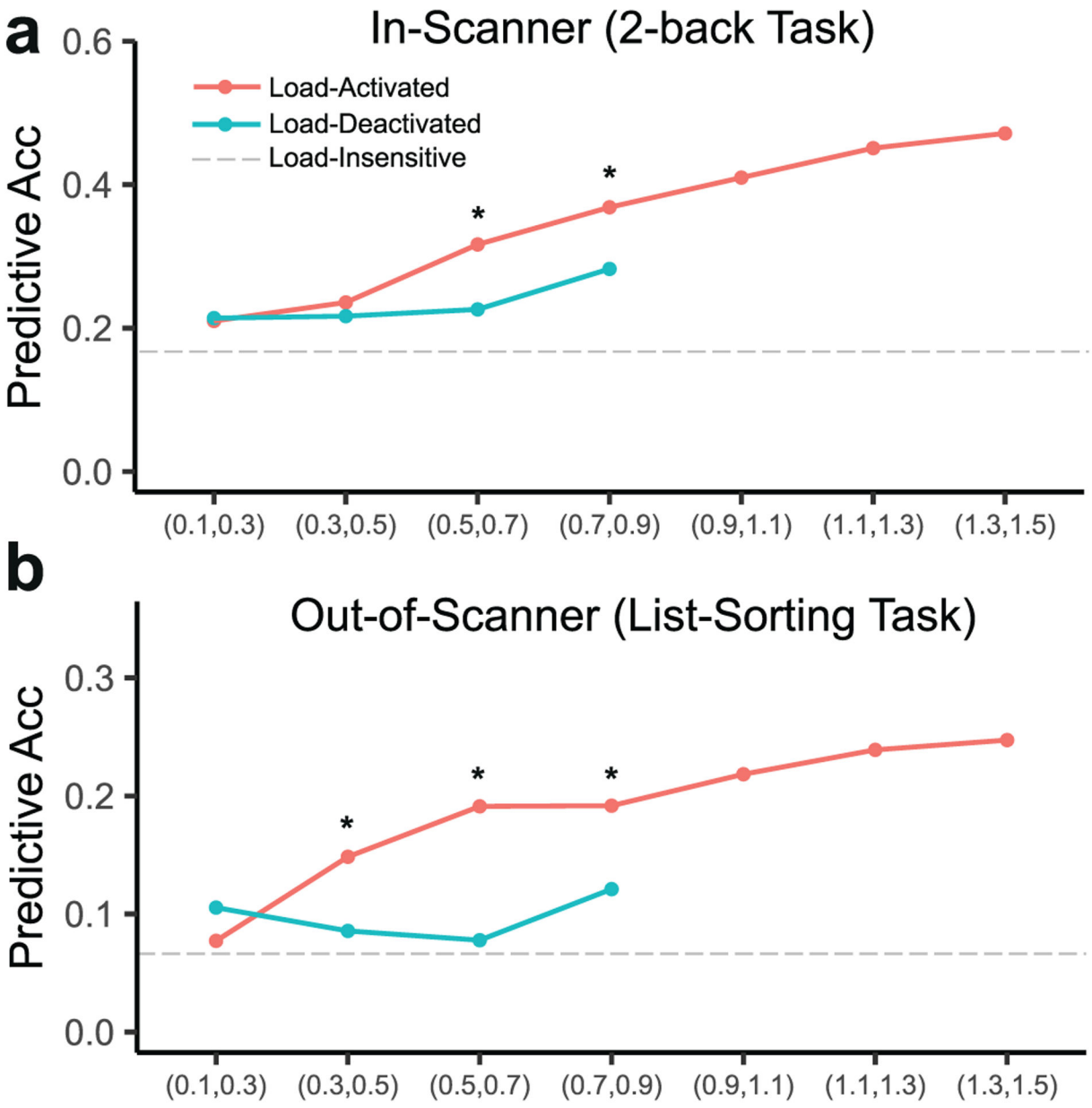

Fig. 5.

The predictive accuracy for each Schaefer parcel bin for a) In-Scanner (N-back) and b) Out-of-Scanner (List-Sorting) performance. Parcels were binned based on load-related effect sizes. Each point represents the predictive accuracy averaged across 100 rounds of sampling from the respective bin. The asterisk indicates a significant difference in predictive accuracy between load-activated and load-deactivated parcel bins being matched in load-related effect size ($p < 0.05$ permutation test). The gray dashed line indicates the predictive accuracy for the load-insensitive bin.

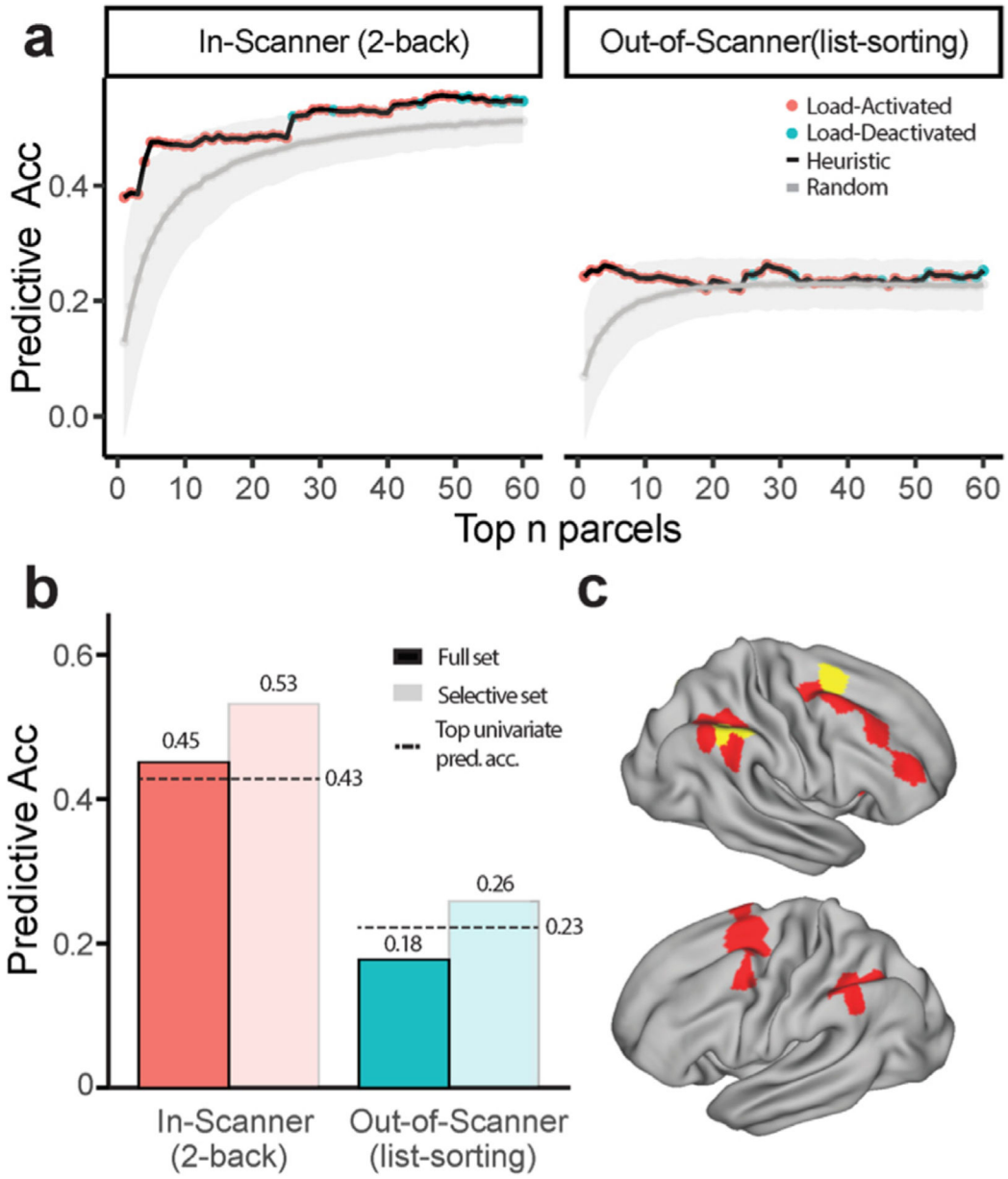

Fig. 6.

The change in model predictive accuracies as adding in the next useful predictive feature. (**a**) The model predictive accuracy when using different number of predictive features. The black line represents the observed predictive accuracy that uses load-related effect sizes to guide feature selection. The color of the data point represents whether the most recently added predictive feature is a load-activated or load-deactivated parcel. The gray ribbon represents the 5% – 95% envelope of the predictive accuracy distribution if the same number of predictive features were randomly sampled. (**b**) The predictive accuracies of models with vs. without feature selection. The bar plot shows the averaged predictive accuracy from the 10-folds cross-validation framework. The horizontal dashed line in the bar plots indicate the

highest univariate predictive accuracies. (**c**) Parcels that show the top 30 largest load effect size, with the yellow color highlighting the top 5 parcels ranked by load-effect size.

Table 1

The spatial relationship between Schaefer parcels being more senstive to within-subject effect and those being more sensitive to between-subjects effect.

Lobe	Within-subject Variation Biased Region Count	Between-subject Variation Biased Region Count
LH_Frontal	16	19
RH_Frontal	17	19
LH_Parietal	10	7
RH_Parietal	9	9

Table 2

The network distribution of Schaefer parcels exhibiting greater sensiviity to within-subject variation (load-effect; left column) and between-subjects variation (individual differences; right column). Networks highlighted in bold exhibited reliable biases in their distribution, as identified through chi-square tests.

Network	Within-subject Variation Biased Region Count	Between-subject Variation Biased Region Count
Control	29	23
Default	43	48
Dorsal Attention	**5**	**41**
Limbic	10	16
Salience/Ventral Attention	21	26
Somatomotor	**62**	**15**
Visual	26	35

Table 3

Identities of the top 30 Schaefer parcels ranked by load-effect size.

Parcel Name	Schaefer ID	Load-related effect size	Univariate predictive accuracy for 2-back performance
RH_Cont_Par_4	335	1.7471310	0.3798165
RH_Cont_Par_5	336	1.7260637	0.3330955
RH_Cont_Par_2	333	1.6731918	0.2973375
RH_Cont_PFCl_15	355	1.6190681	0.4071900
RH_DorsAttn_Post_15	285	1.6166858	0.4406975
LH_DorsAttn_FEF_2	87	1.5648624	0.3713035
LH_Cont_Par_5	131	1.5524267	0.2917752
LH_Cont_Par_6	132	1.5036550	0.3509635
RH_Cont_PFCl_14	354	1.4859233	0.3557810
RH_Cont_Par_6	337	1.4522372	0.3440561
RH_Cont_PFCl_11	351	1.4291677	0.3358864
LH_Cont_pCun_2	145	1.3933306	0.3979757
LH_DorsAttn_Post_9	77	1.3506103	0.3644828
LH_Cont_Par_4	130	1.3391584	0.3259865
RH_Cont_Par_1	332	1.3305392	0.2061880
LH_Cont_PFCmp_1	148	1.3057180	0.3323264
RH_DorsAttn_FEF_2	291	1.3020473	0.3780573
RH_Cont_PFCmp_2	361	1.2908008	0.3191705
RH_DorsAttn_Post_8	278	1.2880328	0.3397741
RH_DorsAttn_Post_11	281	1.2612610	0.2912718
RH_Cont_pCun_2	357	1.2375181	0.3851888
LH_DorsAttn_FEF_4	89	1.2152265	0.3120591
LH_Cont_Par_3	129	1.2085595	0.1586080
LH_Default_PFC_20	185	1.1742447	0.2978756
RH_SalVentAttn_FrOperIns_5	306	1.1611603	0.2683741
LH_Default_pCunPCC_2	191	−1.1601999	0.2161758
RH_SalVentAttn_TempOccPar_7	300	1.1586735	0.1943499
RH_Cont_PFCl_5	345	1.1579546	0.2279562

Parcel Name	Schaefer ID	Load-related effect size	Univariate predictive accuracy for 2-back performance
LH_Default_PFC_24	189	1.1576329	0.3743460
LH_Default_pCunPCC_10	199	1.1422137	0.2523239