How robust are fMRI and EEG data to alternative specifications in representational similarity analyses?

Satwick Sen Sarma (satwick22@iitk.ac.in)
Dept of Cognitive Science, IIT Kanpur, India

Gouravmoy Boruah (gouravmoy22@iitk.ac.in)
Dept of Cognitive Science, IIT Kanpur, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Depts of Cognitive Science and Computer Science, IIT Kanpur, India

Abstract

Computational neuromodeling methods for evaluating representational dynamics involve intricate analysis choices at every stage of the analysis pipeline. Analysis choices for data processing pipelines are generally chosen based upon end to end accuracy metrics and corresponding performance metrics. Psychology research has recently begun to acknowledge the importance of controlling for potential bias introduced by degrees of freedom in data analysis, with specification curve analysis introduced as a principled method for correcting for such biases. In this paper, we conduct a specification curve analysis (SCA) for representational similarity analysis pipelines reported in the literature for fMRI and EEG datasets, respectively. We show that EEG-based RSA analyses are relatively robust to alternative specifications but that fMRI-based analyses are not. Using a novel decision-tree analysis to supplement SCA, we present a potentially more robust pipeline for such analyses.

Keywords: specification curve analysis, fMRI, EEG, representational similarity analysis

Introduction

About ten years ago, (Simmons, Nelson, & Simonsohn, 2011) elegantly demonstrated that, with a small number of *post hoc* analysis decisions, empirical results that satisfy conventional statistics-based thresholds for scientific evidence could be produced for nearly any study. Widespread concerns about *p-hacking* - when analysis pipelines are modified *post hoc* after data has been collected in search of statistically significant results to report - have since led to widespread changes in the way research is reported, e.g. the use of pre-registration and registered reports, more openness to publishing null results or failed replications, and the use of alternatives to null hypothesis testing (Nelson, Simmons, & Simonsohn, 2018).

While the general principle that data-dependent analysis offers a garden of forking paths to false positive results has been widely accepted (Gelman & Loken, 2016), principled solutions have been hard to come by. Recently, specification curve analysis (SCA) has emerged as an interesting method to assess the robustness of specific results, given access to data (Simonsohn, Simmons, & Nelson, 2020). The basic idea behind SCA is blindingly simple - authors' journey from data to results is chaperoned by a series of decisions about how to conduct analyses. At each point in this series, an alternative researcher may presumably have selected a different choice than the one the authors chose. Therefore, a combinatorially large set of possible *specifications* exists to analyze the data, of which the original *specification* that led to the claimed result is one. SCA permits one to redo the analysis across this set and report whether the result continues to be statistically significant across most reasonable specifications of the analysis chain or whether a very specific set of choices leads to a statistically significant result, while most others do not (Simonsohn et al., 2020). A result that falls in the former category may be more reliable than a result that falls in the latter.

The past decade has also seen an efflorescence of research adopting the framework of representational similarity analysis (RSA) to try to find results that unify brain-activity data, behavioral data, and computational models (Kriegeskorte, Mur, & Bandettini, 2008). RSA makes strong assumptions about the nature of representations, viz. that brain or behavioral activity seen in an experimental condition can be directly treated as a stimulus-condition representation and then proceeds to use correlations between all possible condition pairs to create entries in a representational dissimilarity matrix (RDM). Comparing RDMs across modalities, researchers seek to identify models that are strongly correlated with data-based RDMs.

However, in practice, data analysis pipelines for estimating brain activity from raw hemodynamic responses are saturated with a diversity of specification choices. For example, to estimate feature vectors based upon linearizing encoding models (Naselaris, Kay, Nishimoto, & Gallant, 2011) from BOLD fMRI, the estimation of best-fitting canonical hemodynamic response functions can be performed in a multitude of ways (Pedregosa, Eickenberg, Ciuciu, Thirion, & Gramfort, 2015). Estimation strategies using voxel-wise best-fitting HRF often introduce biases due to idiosyncratic noise embedded in the signal in individual voxels. Estimating HRFs for entire regions of interest(ROI), on the other hand, results in more unbiased, but also more variable, estimates (Pedregosa et al., 2015).

In this paper, we try to understand the robustness of RDMs acquired from fMRI data using a specification curve analysis. In contrast with the large garden of forking paths relevant for fMRI data analysis, EEG data analysis in RSA offers relatively fewer degrees of flexibility. As a baseline for comparison with the fMRI analysis, we also present a similar specification curve analysis for EEG-based RDMs. Finally, we present a decision-tree based search methodology for identi-

fying the most robust specification from a set of of alternative specifications, and specifically identify such a specification for the representational similarity analysis of fMRI data.

Methods

Datasets

We used the THINGS-data fMRI dataset (Hebart et al., 2023) comprising 720 object concepts from the THINGS concept class (Hebart et al., 2019). THINGS-data contained MRI data for stimulus presentation of 12 exemplars of each of the 720 object categories over 12 sessions for each subject. For our analysis,we used the data of a randomly sampled subject from the pool of three subjects. Stimulus was presented in an event related design for 500ms, followed by fixation of 4s. Furthermore, 6 category selective functional runs for faces, body parts, scenes, words, and objects were also recorded for each subject. MRI recordings across multiple sessions on diverse sets of object categories posited the optimal basis for reliability assessment of multivariate analyses of single trial responses in evoked BOLD responses.

For EEG, we used the THINGS-EEG dataset (Grootswagers, Zhou, Robinson, Hebart, & Carlson, 2022) for our analysis. We randomly pooled 5 subjects from a sample size of 50 subjects. The data used in the analysis was for the 200 validation images presented randomly in a rapid serial visual presentation paradigm (RSVP), repeated across 12 sequences. Individual images were presented for 50 ms, followed by a blank screen, which lasted for another 50 ms. The same 200 images were repeated for all the subjects; the uniformity of the stimuli across participants is vital for an accurate reliability assessment of the underlying representational dynamics.

Data Preprocessing

fMRI Functional images for each run were preprocessed by performing slice timing correction, rigid head motion correction, field map-based susceptibility distortion correction, alignment of functional space to individual subject's T1-weighted anatomical template, brain tissue segmentation, and pial and white matter surface reconstruction. ICA was performed post additional preprocessing of each functional run with spatial smoothing and high pass filtering. MELODIC ICA (Beckmann & Smith, 2004) was implemented for component estimation, and correlation coefficient thresholds as specified in THINGS-data (Hebart et al., 2023) were followed for noise component detection.

EEG The data (Grootswagers et al., 2022), comprising 64-channel EEG recording, was preprocessed using EEGLab (Delorme & Makeig, 2004). Data was filtered by applying a Hamming-windowed FIR filter with 0.1 Hz highpass and 100 Hz lowpass filters. The electrodes were re-referenced to average, and the data was downsampled to 250 Hz. The continuous EEG data was epoched into trials ranging from 100 ms before the stimulus onset to 1000 ms after stimulus onset.

fMRI β-weights estimation

Original Specification The original specification for data analysis of fMRI data was taken from (Hebart et al., 2023). Following their recipe, BOLD time series for each run was normalized using percent signal change ("The General Linear Model", 2011) to eliminate the effects of variable scale and the resulting effect size variability over each individual voxel. Time course normalization was performed at two levels, viz. at runwise signal level and at global signal level across all runs over the sessions, to further negate any effect of signal level differences across sessions. For the global level, we performed normalization by removing the mean over all runs of the signal time series for every voxel. Although time-course normalization is not essential for GLM estimation, for comparing voxel activity spaces using representational dissimilarity matrices (RDM), the scale of signals at a coherent level allows for more reliable comparisons across sessions. Normalized functional runs were noise normalized by fitting GLMs runwise with noise components produced by MELODIC ICA (Beckmann & Smith, 2004) and drift components estimated by fMRIprep. For each run, residuals were calculated after model fit and z-standardized for downstream analysis. We estimated β weights for single trials using a similar analysis pipeline as mentioned in THINGs-data (Hebart et al., 2023). We changed some parameter estimation strategies to reduce bias in the resulting analysis due to confounds. Firstly, for finding the best hemodynamic response function (HRF), we estimated the best-fit HRFs for each ROI based on the mean R^2 over all sessions from a set of 20 canonical HRF functions (Allen et al., 2022). Finding specific HRF for each ROIs takes into account the variability of these basis response functions across regions (Handwerker, Ollinger, & D'Esposito, 2004; Badillo, Vincent, & Ciuciu, 2013) while also mitigating the chances of overfitting to noise when voxel-specific HRFs are used. We also fixed the regularization parameter $\alpha = 0.1$ for all voxels as our initial assessment displayed non-existent variability after an exhaustive search over the hyperparameter space for α ranging from 0.1 to 0.9. The estimation of amplitude, i.e., β- weights from convolved trail design matrices with hemodynamic basis functions, was performed by fitting a ridge-regression model for each voxel.

Alternative Specifications Devising alternative specifications for single trial response estimates in fMRI included developing alternative strategies for time course normalization (Liu, Glover, Mueller, Greve, & Brown, 2013; Yan, Craddock, Zuo, Zang, & Milham, 2013) of the signals and an alternative best-fit HRF estimation method. In the original specification, time course normalization was performed sequentially, voxel-wise. Firstly, runwise by utilizing percent signal change(PSC) then subsequently using baseline correction using the global mean over all sessions and runs for the voxel. Alternatively, standardizing levels of signals(RTN) can be performed by z-standardization(Z-S) (Zhao et al., 2018) or baseline correction for the mean(CFM) (de Beeck,

2010) for each run. Similarly, global normalization(GTN) can be performed by z-standardization(RZ-S), re-centering from mean(RCFM) or percent signal change(RPSC), and even just using the runwise normalized signal (OF) by itself since the comparison across sessions is based on amplitude estimates of linear encoding models, which is independent of signal scale or level variances across sessions. In total, this set of combinations resulted in $3*4=12$ specifications. Choosing whether to standardize residuals(ZSR) or not(NZSR) after performing ICA denoising created two further specifications, producing $12*2=24$ specification alternatives for downstream processing pipeline. Further downstream specification choices were devised by utilizing varying hemodynamic response function(HRF) fitting strategies and variations in ROI estimation process. HRF fitting was performed for each region of interest (ROIs) (Ciuciu, Idier, Roche, & Pallier, 2004) individually and was performed by estimating best fitting HRFs in two alternative strategies. One approach selects the best-fitting HRF(BHRF) for each ROI over all sessions cumulatively(H) and the other selects the most frequent best-fit HRF(HM) across sessions. Both specifications used the best-fit metric estimated from the mean R^2 over the voxels (Allen et al., 2022) comprising the ROI. For estimating β-weights, we developed $12*2*2=48$ alternative specification, one of which is the original specification itself.

EEG Decoding Analysis

Original Specification. The original specification for data analysis of EEG data was taken from (Grootswagers et al., 2022). Following their approach, our multivariate decoding analyses using EEG-evoked responses for 200 images across n=5 subjects was performed within subjects, while subsequent analysis was done at the group level. We then constructed RDMs (Kriegeskorte et al., 2008) that map out the dissimilarity patterns evoked by each stimulus pair, a total of $\binom{200}{2}$, i.e., 19,900 pairs. The pipeline was used for all stimulus-present epochs[-100 to 1000ms].

In the decoding analyses, the voltages of all 64 EEG channels were used as features for each time point. A regularized (λ=0.01) linear discriminant classifier was trained using a leave-one-sequence-out cross-validation procedure. This involved reserving one image presentation sequence from each category as test data while training the classifier on the remaining image presentation sequences. This resulted in (19,900 image condition pairs x 275 eeg time points) shaped EEG-RDM containing the mean classification accuracy scores for the image pairs in the left-out sequences for each subject. The RDMs were then averaged across the image pairs to calculate the mean pairwise classification accuracy over time.

Alternative Specification Alternative specifications were devised by performing the leave-k-sequence out cross-validation procedure with varying k-values and training a regularized (λ=0.01) linear discriminant classifier for image pair classification. Here, k sequences, ranging from 1 to 10, were reserved as test data while training the classifier on remaining (12 - k) image presentation sequences for each category. This pipeline was used for all stimulus-present epochs [-100 to 1000 ms]. The mean classification accuracy for the image pairs in the left-out sequences was calculated and recorded in the RDMs for each k-value for all the subjects, which were then averaged across the range of k-values to calculate the mean pairwise classification accuracy over time for each k. In total, 10 specifications, one for each k-value, were devised for the EEG decoding pipeline.

fMRI Region of Interest (ROI) Map Estimation

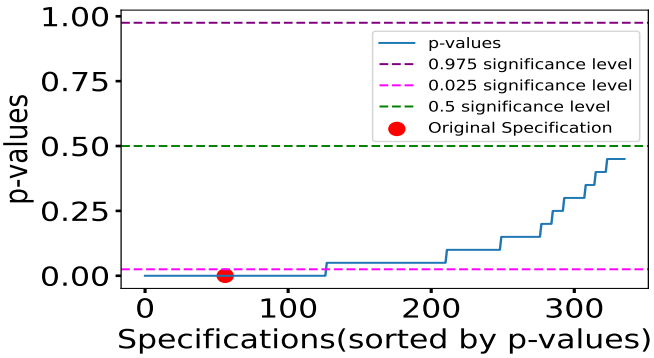
Original Specification This method also followed the approach used in (Hebart et al., 2023). THINGS-data (Hebart et al., 2023) included six functional imaging for category localizer experimental runs for each participant. Based on the preference of activation pattern of each region of interest to particular object categories, we defined T-contrast schema for estimating ROIs (Poldrack, 2007). Statistical parametric threshold maps were aggregated across spatially smoothed six functional runs with a fixed effects model (Woolrich, Behrens, Beckmann, Jenkinson, & Smith, 2004) and corrected for multiple comparisons $(\text{cluster threshold} = 0.0001, \text{extent threshold} = 3.7)$.Resulting maps were intersected with existing category-selective brain area parcels (Julian, Fedorenko, Webster, & Kanwisher, 2012) to account for single subject noise sensitivity, thus producing the final estimates for EBA, FFA, LOC, and PPA regions.

Alternative Specifications For estimating the contrast maps, we sampled an exhaustive set of 3,4 and 5 runs, respectively, from the set of 6 totally available runs(RU) resulting in $\binom{6}{k}$ maps for $k=3,4,5$. For each of the sets of 3(3R), 4(4R), and 5(5R) functional runs, we created two masks from each set of threshold maps(MM) — by union(U) and intersection(I) of the maps. In total, we estimated 7 maps, including the original map from 6 runs(6R).

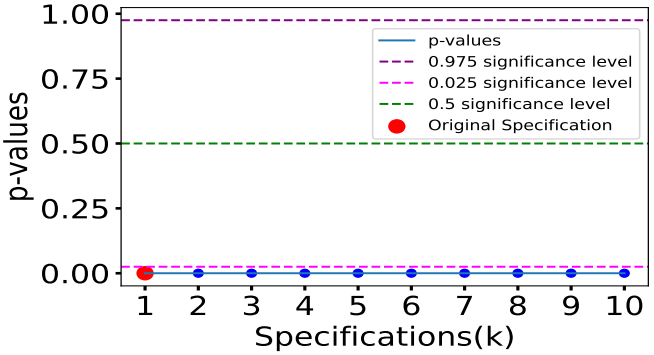
Combining both alternatives at the stage of calculating beta-weights and at the stage of calculating the ROI map, our alternative specification choices yielded a set of $48*7=336$ specifications, including the original specification. The set of analysis forks, with the possible choices available at each one, is listed in Table 1.

Table 1: Summary Of alternative specifications for fMRI analysis. Each column lists the possible options within each analysis fork (column header). 1^{st} row = original spec.

RTN	GTN	SR	BHRF	RU	MM
PSC	RCFM	ZSR	H	6R	I
Z-S	RPSC	NZSR	HM	3R	U
CFM	RZ-S			4R	
	OF			5R	



(a) SCA for fMRI



(b) SCA for EEG

Figure 1: **p-values** for **fMRI** and **EEG** specifications from **SCA**. For fMRI, specifications are arbitrarily ordered based on p-values. For EEG, specifications correspond to the value of k in leave-k-sequence out cross-validation.

Specifications Curve Analysis

In order to perform Specifications Curve Analysis (SCA) (Simonsohn et al., 2020), we generated the null distribution by shuffling the stimulus labels for the conditions for both fMRI and EEG recordings. For fMRI, we shuffled the labels of the stimulus over each session and estimated the average RDM over each session by computing the mean RDM of individual shuffled session-wise RDMs. To calculate the similarity index across the specifications, we calculated Kendall's τ_A (Khaligh-Razavi & Kriegeskorte, 2014; Nili et al., 2014) for each specification against the original specification using the resultant RDMs from the 19 shuffled samples, which were 19 in total. Based on Kendall's τ_A scores of each specification for 19 shuffles and the original sample, we estimated p-values for assessing the significance of the specifications. For EEG, we performed random shuffling, 10 in total, of the stimulus labels of the time-series data for each stimulus presentation and constructed the null distribution. We then calculated the subject-wise RDMs and estimated the mean time-varying decoding accuracy for each shuffle. The p-values were calculated using Kendall's τ_A mean time-varying decoding accuracy correlation for each k-value pairing from the original sample against the decoding accuracy of the k-value pairings of the 10 shuffled samples.

Results

We present two sets of results. First, we present the reliability of fMRI activation coefficients and EEG decoding analysis for specification changes. Second, we show how to identify the most robust specification for data analysis possible in the set of possible fMRI analysis specifications we have defined.

Robustness to alternative specifications

fMRI We performed a two-sided significance test for the rank correlation, Kendall's τ_A, estimates of each specification for the original data using p-value estimation (see Methods) from the simulated null sample with 19 shuffles.

Table 2: Specification choices producing the largest number of significant correlations.

RTN	GTN	SR	BHRF	RU	MM	#
any	OF	any	any	3R	U	12
any	RCFM	any	any	5R	I	12
any	RZ-S	any	any	5R	I	12
any	OF	any	any	5R	I	12
any	PSC	any	H	5R	U	6
any	RZ-S	ZSR	any	3R	U	6

Table 3: Specification choices producing the largest number of non-significant correlations.

RTN	GTN	SR	BHRF	RU	MM	#
any	any	any	any	3R	I	48
CFN	any	any	HM	4R	any	16
PSC	any	any	HM	4R	any	16
any	OF	any	any	5R	U	12
any	PSC	any	any	5R	I	12
Z	RCFM	ZSR	HM	4R	any	8

We found significance [$p < 0.025$] for only 126 alternative specifications out of the total 335 alternative specifications in addition to the original specification.

The specification significance curve is illustrated in Figure 1, showing that for the other set of 209 specifications, the correlation between the RDM estimated using the particular alternative specification with the RDM produced using the original specification is not significantly different from estimated correlations between specifications in the null samples.

We also found that the original specification is quite fragile with respect to changes in analysis choices. Figure 2(top) plots the fraction of alternative specifications yielding non-significant correlations when a single analysis decision (of the six we evaluated) is changed in the original analysis specification. Of the total 11 specifications achievable by a single

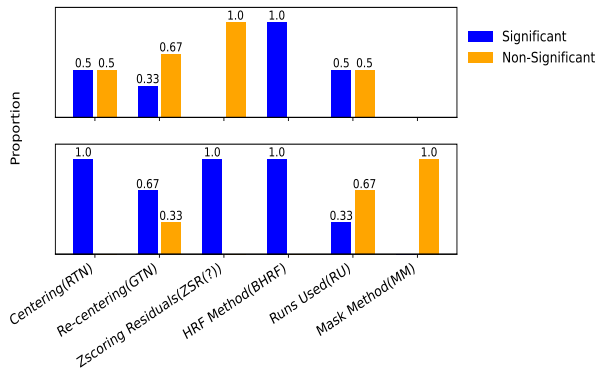


Figure 2: Proportion in specification shifts caused by a change in one analysis choice. The top panel shows the shift for the original specification. The bottom panel shows the same for the observed most robust specification.

perturbation of the analysis method, only 6 yield significantly different correlations from the null distribution.

EEG Kendall's τ_A was calculated between the mean time-varying decoding accuracy of the original specification against the mean time-varying decoding accuracies of the 10 alternative specifications. Similarly, for each shuffle, we also calculated Kendall's τ_A between the decoding accuracy of the shuffled original specification and the 10 shuffled alternative specifications to estimate the null. Then, a significance test was performed for the rank correlation relations, Kendall's τ_A, for specifications for the original data using the p-value estimation approach based on a simulated null sample through 10 shuffles. The analysis showed significant [$p < 0.025$] results for all the 10 specifications, as seen in Fig 1(b). Unlike fMRI, which showed robustness for 126 alternative specifications out of 335, the alternative specifications curve analysis for EEG indicates robustness for all the specification choices of the pipeline. The decoding model consistently produced significant results regardless of the number of left-out (k) sequences, indicating model stability across cross-validation configurations. Hence, the model reliably captures underlying neural patterns parallel to the stimuli shown, and its ability to decode the EEG time series data is not a result of overfitting to a particular sequence of data but is effective across various splits, indicating model generalizability. This shows that the EEG decoding pipelines are reliable tools for analyzing the dynamics of neural representation.

Finding robust fMRI specifications

The set of alternative specifications defined in SCA is obtained by a branching process in the garden of analysis forks, which eventually yields specifications that are significantly or non-significantly different from the null distribution. Therefore, it is possible to model the mapping of the specifications to the binary prospect of their being significant or not by fitting a decision tree classifier (Quinlan, 1983) using the specification choices as features and significance as the binary target variable.

Examination of the structure of the learned tree reveals the importance of each parameter specification based on frequency. Decision tree paths leading to leaf nodes with the most number of significant and non-significant specifications are tabulated in Tables 2 and 3 respectively. Evidently, using 5 runs seems to be most predictive of significance for specifications as shown in Table 2, though even here, selecting some choices in other analysis forks can lead to clusters of non-significant specifications, as shown in Table 3.

Based on the decision tree's structure, we defined the most robust specification as the one that leads to the least number of non-significant specifications if one analysis choice is varied. For the set of alternative specifications we used, specification shifts for the most robust specification (PSC-OF-ZSR-H-5R-I) are shown in Figure 2(bottom). For this specification, changes in either HRF estimation strategy, z-standardization of residuals, or normalization methods do not affect the significance of the correlations. Only using union instead of intersection for ROI definition results in a non-significant combination.

In comparison with the original specification, this one chooses to not use any form of global normalization. Thus, simplifying the original specification in this way, it is possible to arrive at one that is substantially more robust to further alternative specifications.

Discussion

Cognitive science is deluged with fMRI-based studies using stimulus-yoked designs to make claims about localization of various cognitive phenomena (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). In view of the large garden of forking paths known to exist in the processing of fMRI data for stimulus-linked experiments, we conducted a specifications curve analysis for fMRI data-based representation similarity analyses (Simonsohn et al., 2020). We showed that, in comparison with a baseline EEG-based representation similarity analysis pipeline, the fMRI pipeline shows significant variability in the outcomes of multiple theoretically reasonable alternative specifications. In particular, only a third of the set of alternative specifications achieved statistical significance in our analysis. We then showed, using a novel decision-tree based approach, that a small change in the original specification could make it more robust.

We note though, that even the most robust specification identified in our exercise would still not reduce the overall fragility of the fMRI analysis pipeline. For instance, switching the mask estimation method from I to U in the robust specification discovered in our analysis yields a large number of non-significant alternatives, as evident by the presence of this modified specification in Table 2. Moreover, we could not identify any theoretically sensible pattern to characterize specifications failing to pass the significance test based on SCA. Thus, we inevitably conclude that, in addition to its known unreliability (Elliott et al., 2020), single trial fMRI

data is also fragile to analysis choices in its information processing pipeline. Therefore, caution is warranted in interpreting representational similarity results using fMRI data.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Badillo, S., Vincent, T., & Ciuciu, P. (2013). Group-level impacts of within-and between-subject hemodynamic variability in fmri. *Neuroimage*, *82*, 433–448.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, *23*(2), 137–152.

Ciuciu, P., Idier, J., Roche, A., & Pallier, C. (2004). Outlier detection for robust region-based estimation of the hemodynamic response function in event-related fmri. In *2004 2nd ieee international symposium on biomedical imaging: Nano to macro (ieee cat no. 04ex821)* (p. 392-395 Vol. 1). doi: 10.1109/ISBI.2004.1398557

de Beeck, H. P. O. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fmri analyses? *Neuroimage*, *49*(3), 1943–1948.

Delorme, A., & Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9–21.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., ... Hariri, A. R. (2020). What is the test-retest reliability of common task-functional mri measures? new empirical evidence and a meta-analysis. *Psychological science*, *31*(7), 792–806.

Gelman, A., & Loken, E. (2016). The statistical crisis in science. *The best writing on mathematics (Pitici M, ed)*, 305–318.

The General Linear Model. (2011, 03). In *Statistical Analysis of fMRI Data.* The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/8764.003.0007 doi: 10.7551/mitpress/8764.003.0007

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., & Carlson, T. A. (2022). Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, *9*(1), 3.

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, *21*(4), 1639–1651.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., ... Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, *12*, e82580.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357–2364.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Liu, T. T., Glover, G. H., Mueller, B. A., Greve, D. N., & Brown, G. G. (2013). An introduction to normalization and calibration methods in functional mri. *Psychometrika*, *78*, 308–321.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology*, *69*, 511–534.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven hrf estimation for encoding and decoding models. *NeuroImage*, *104*, 209–220.

Poldrack, R. A. (2007). Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, *2*(1), 67–70.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In *Machine learning* (pp. 463–482). Elsevier.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, *20*(3), 470–477.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for fmri group analysis using bayesian inference. *Neuroimage*, *21*(4), 1732–1747.

Yan, C.-G., Craddock, R. C., Zuo, X.-N., Zang, Y.-F., & Milham, M. P. (2013). Standardizing the intrinsic brain: to-

wards robust measurement of inter-individual variation in 1000 functional connectomes. *Neuroimage*, *80*, 246–262.

Zhao, N., Yuan, L.-X., Jia, X.-Z., Zhou, X.-F., Deng, X.-P., He, H.-J., … Zang, Y.-F. (2018). Intra-and inter-scanner reliability of voxel-wise whole-brain analytic metrics for resting state fmri. *Frontiers in neuroinformatics*, *12*, 54.