Stimulus repetition and sample size considerations in item-level representational similarity analysis

Stephen Mazurchuk, Lisa L. Conant, Jia-Qing Tong, Jeffrey R. Binder & Leonardo Fernandino

View supplementary material

Published online: 10 Jul 2023.

Submit your article to this journal

View related articles

View Crossmark data

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Check for updates

Stimulus repetition and sample size considerations in item-level representational similarity analysis

Stephen Mazurchuk [a,b], Lisa L. Conant[a], Jia-Qing Tong[a,b], Jeffrey R. Binder[a,b] and Leonardo Fernandino[a,c]

[a]Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA; [b]Department of Biophysics, Medical College of Wisconsin, Milwaukee, WI, USA; [c]Department of Biomedical Engineering, Medical College of Wisconsin and Marquette University, Milwaukee, WI, USA

ABSTRACT

In studies using representational similarity analysis (RSA) of fMRI data, the reliability of the neural representational dissimilarity matrix (RDM) is a limiting factor in the ability to detect neural correlates of a model. A common strategy for boosting neural RDM reliability is to employ repeated presentations of the stimulus set across imaging runs or sessions. However, little is known about how the benefits of stimulus repetition are affected by repetition suppression, or how they compare with the benefits of increasing the number of participants. We examined the effects of these design parameters in two large data sets where participants performed a semantic decision task on visually presented words. We found that reliability gains from stimulus repetition were strongly affected by repetition suppression, both within and across scanning sessions separated by multiple weeks. The results provide new insights into these experimental design choices, particularly for item-level RSA studies of semantic cognition.

Introduction

Representational similarity analysis (RSA) is an increasingly popular multi-voxel pattern analysis method in which measured neural activity patterns are used to construct a neural representational dissimilarity matrix (RDM), which is then compared to dissimilarity matrices predicted by one or more stimulus models or measured using an entirely different neural encoding technique (Kriegeskorte et al., 2008). Common aims of RSA include comparing different stimulus models in terms of strength of correlation with the neural RDM as a means of adjudicating the validity of different models, and mapping the brain regions where neural activity patterns reflect a particular kind of information. Achieving these aims is likely to depend on the reliability of the estimated neural RDM, which like all physiological recordings is affected by noise and individual variability. Here we consider two strategies for enhancing reliability of the neural RDM, namely, presenting the stimulus set multiple times (and combining the signal across presentations) and increasing the number of participants.

Effects of sample size on reliability have been a focus of some univariate fMRI studies (e.g.Thirion et al., 2007; Turner et al., 2018). These studies demonstrate important trade-offs between the number of participants and amount of data collected in each participant.

When limited amounts of individual-level data are acquired, results may still not be replicable even with large sample sizes (Nee, 2019; Turner et al., 2018). Increasing the amount of data acquired for each participant may allow adequate reliability to be achieved with smaller sample sizes (Nee, 2019). However, considerably less is known about how these factors impact the reliability of results in multivariate (e.g. multivoxel) pattern analysis.

Except for the choice of distance metric (Allefeld & Haynes, 2014; Bobadilla-Suarez et al., 2020; Ritchie et al., 2021; Walther et al., 2016), little attention has been given to the factors that may contribute to the reliability of the neural RDM. In most RSA studies, stimuli are repeated so that the signal-to-noise ratio can be increased through averaging. This is particularly true for studies in which the RDM is constructed from neural responses to single items (rather than categories of items), which we refer to as "item-level RSA". However, the benefit of averaging across repetitions may be offset by the possibility of response adaptation, also known as repetition suppression, in which subsequent neural responses to the same stimulus are weaker as measured by hemodynamic methods (Dobbins et al., 2004; Grill-Spector et al., 2006; Lee et al., 2020). Notably, it is unclear whether this decreased

---

magnitude of responses is problematic for RSA, as Arbuckle et al. (2019) found that representational geometry stayed relatively stable across significant changes in average fMRI activity. While repetition suppression has been extensively investigated in univariate analysis of fMRI data, its impact on RSA results is largely unknown.

Exemplifying the need for a better understanding of these trade-offs, a search of the RSA literature restricted to item-level RSA studies using visually presented single words as stimuli reveals a large variation in participant numbers and stimulus presentations (Supplementary Table 1). Many studies have around 20 participants (Carota et al., 2021; Dong et al., 2021; Li et al., 2019; Meersmans et al., 2022; Staples & Graves, 2020; Wang et al., 2018), with a range from 9 (Anderson et al., 2016) to 51 (Guo et al., 2022). The median number of stimulus presentations was 4 (Borghesani et al., 2016; Bruffaerts et al., 2013; Liu et al., 2023; Meersmans et al., 2022), with a range from 1 (i.e. no repetitions) (Dong et al., 2021; Gao et al., 2022; Guo et al., 2022; Li et al., 2023; Staples & Graves, 2020) to 12 (Fischer-Baum et al., 2017).

While the effects of stimulus repetition and participants sample size on neural RDM reliability are largely unexamined, the importance of between-subject reliability for the interpretation of RSA results is widely recognised (Ritchie et al., 2021). Reliability of the neural RDM is an essential prerequisite for making valid inferences regarding relationships with model RDMs and discriminating among models. In the RSA literature, the reliability metrics typically used are the upper and lower bound estimates of the noise ceiling, which are measures of the highest potential performance of a model for a given data set. Noise ceiling estimates allow the obtained RSA correlation to be interpreted in the context of the limitations of the specific experiment (e.g. level of measurement noise, sample size, amount of individual-level data). In the current study, to further assess interindividual consistency, or internal consistency of the group-averaged neural RDM, and to relate our analysis to the broader context of classical measurement theory, we also examined another measure of reliability, Cronbach's alpha (hereafter referred to as "alpha"). This choice was motivated by the close connections between alpha and other reliability measures such as split-half correlation and intra-class correlation (ICC), which allow it to also be interpreted as a measure of reproducibility (Bravo & Potvin, 1991), along with its quick and efficient calculation, allowing for resampling analysis across varying numbers of participants and stimulus presentations. In addition, alpha allows researchers to estimate

reliabilities for sample sizes larger than the obtained sample (de Vet et al., 2017).

Here we examined sample size and stimulus repetition effects in two large data sets using item-level RSA with lexical stimuli. In both studies, a large stimulus set was presented 6 times over three scanning sessions on separate days. The analyses focused on a region of interest (ROI) encompassing primarily multimodal cortical areas in the frontal, temporal, and parietal lobes, derived from a previous ALE meta-analysis of semantic cognition studies (Binder et al., 2009). This large ROI provides a global estimate of RDM reliability across the high-level association areas of the cortex, averaging over local differences in reliability due to spatial variation in MRI signal quality. To ensure that the pattern of results was not dependent on the particular choice of ROI, we also replicated some of the analyses in smaller ROIs defined within a commonly used cortical parcellation (Desikan et al., 2006). The goal of the analyses is to provide researchers with useful information on the potential trade-offs between the number and schedule of stimulus presentations and the number of participants to help optimise study design and use of limited resources, such as scanner time. Importantly, the intention is not to make strong recommendations regarding optimal values for these parameters, as these can vary based on the experimental design, but rather to examine how specific design choices may affect reliability in condition-rich, item-level RSA studies in terms of their general trends.

Materials and methods

Participants

Study 1 included 40 adult, native speakers of English (25 women) with a mean age of 28 years. Study 2 included 39 adult, native speakers of English (27 women) with a mean age of 27 years. Four individuals participated in both studies. All participants were right-handed according to the Edinburgh Handedness Scale (Oldfield, 1971), had at least a high school education, and had no history of neurologic or psychiatric conditions. All participants provided written informed consent. The studies were approved by the Medical College of Wisconsin Institutional Review Board.

Stimuli and task

The stimuli in Study 1, described in detail in a previous publication (Fernandino et al., 2022), consisted of 320 English nouns, half of which were names of objects and half names of events. Object nouns were animals,

tools, plants/foods, and vehicles; event nouns were sounds, negative events, social events, and communication events. The stimuli in Study 2, described in Tong et al. (2022), were 300 English nouns from 6 categories: animals, artifacts, plants/foods, body parts, human traits, and quantities.

Both experiments used the same stimulus presentation procedure and task. On each trial, a noun was displayed in white font on a black background at the centre of the screen for 500 ms, followed by a 2.5-s blank screen. Participants were instructed to rate each word on how often they encountered the corresponding entity or event in their daily life, using a scale from 1 ("rarely or never") to 3 ("often"). Responses were indicated using three keys operated with the right hand. Each trial was followed by a central fixation cross with variable duration between 1 and 3 s (mean 1.5s). Each run started and ended with an 8-s fixation cross. Stimulus presentation and response recording were performed with Psychopy 3 software (Peirce et al., 2019) running on a Windows desktop computer and a Celeritas fibre optic response system (Psychology Software Tools, Inc.). Stimuli were displayed on an MRI-compatible LCD screen positioned behind the scanner bore and viewed through a mirror attached to the head coil.

In both studies the entire stimulus set was presented six times with a different pseudo-randomized order for each presentation and each participant. The presentation of the complete set of stimuli was split into 4 runs each containing 80 trials for Study 1, and 3 runs each containing 100 trials in Study 2. Data collection was performed over the course of three scanning sessions on separate days, with two complete stimulus set presentations per session. For Study 1 the interval between sessions 1 and 2 averaged 18 days (sd = 24.8), and the interval between sessions 2 and 3 averaged 25 days (sd = 37.1). For Study 2, the respective intervals between the sessions were 23 days (sd = 29.1) and 23.5 days (sd = 26.0). These intervals did not significantly differ in either study (Wilcoxon $p = .98$ and $p = .52$ in studies 1 and 2, respectively; **Figure S1**)

MRI

Scanning was performed on a GE Healthcare Premier 3 T MRI scanner with a 32-channel Nova head coil at the Medical College of Wisconsin's Centre for Imaging Research. Each session consisted of a structural T1-weighted MPRAGE scan, a structural T2-weighted CUBE scan, 3 pairs of T2-weighted spin echo echo-planar scans (5 volumes each) acquired with opposing phase-encoding directions for later image unwarping, and either 8 (Study 1) or 6 (Study 2) gradient-echo echo-

planar imaging (EPI) functional scans (multiband factor = 4, TR = 1500 ms, TE = 33 ms, flip angle = 50, in-plane matrix = 104 × 104, slice thickness = 2.0 mm, axial acquisition, 68 slices, field-of-view = 208 mm, voxel size = 2 × 2 × 2 mm). Studies 1 and 2 had 251 and 311 volumes per run, respectively.

FMRI processing

MR images were preprocessed via a containerised version of the fMRIprep 22.0.1 pipeline (Esteban et al., 2019). Three field maps were used to estimate a B0-non-uniformity map using "topup" (FSL 6.0.5). The three T1-weighted (T1w) images were corrected for intensity non-uniformity (INU). A T1w reference map was computed by averaging these T1w images after registration using "mri_robust_template" (FreeSurfer 7.2.0). The T1w reference was then skull-stripped with a Nipype implementation of the "antsBrainExtraction.sh" workflow. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter, and gray matter was performed on the brain-extracted T1w image using FAST (FSL 6.0.5). Brain surfaces were reconstructed using "recon-all" (FreeSurfer 7.2.0). For each of the EPI runs for each participant, a reference volume and its skull-stripped version were generated using fMRIprep. Head-motion parameters with respect to the EPI reference were estimated before any spatiotemporal filtering using MCFLIRT (FSL 6.0.5). EPI runs were slice-time corrected, and the EPI reference was co-registered to the T1w reference using "bbregister" (FreeSurfer), a boundary-based registration algorithm with six degrees of freedom. Grayordinate files using the fsLR-32k_midthickness cortical surface model (Glasser et al. 2013) containing 91k samples were also generated using the highest resolution fsaverage as an intermediate standardised surface space. No additional spatial smoothing was applied.

Beta and t values were calculated directly on the surface time-series data using 3dREMLFIT in AFNI (Cox, 1996), incorporating 6 motion estimates, their derivatives, CSF and white matter regressors, and 4th order baseline polynomials for detrending. We separately z-scored response time estimates across each set of stimulus presentations and included them as nuisance regressors in regression models having at least 2 presentations of the stimuli. In addition, we censored volumes that had a framewise displacement greater than .9 mm. Each word, including those with multiple presentations, was its own regressor in every model (i.e. except for reaction time, all models had the same number of regressors). We ensured that no functional data from omitted runs contributed to the estimated beta values calculated from

different presentation combinations by running separate regressions for every set of presentation combinations.

ROI selection

The primary ROI used to generate the neural RDMs was a functionally defined mask based on a previous activation likelihood estimate (ALE) meta-analysis of 120 functional neuroimaging studies of semantic language processing (Binder et al., 2009). The ALE map was projected on the HCP fsLR-32k_midthickness cortical surface model and binarized at a significance threshold of $p < .01$. The ROI included substantial portions of association cortex in frontal, parietal, and temporal cortex (Figure 1). We chose to use a pre-defined semantic ROI to minimise bias that could be introduced in the analysis by selecting an ROI through a reliability measure and subsequently performing reliability analysis. However, to ensure the generality of our results, we repeated the analyses in 5 anatomically defined ROIs associated with language processing. This was done by projecting the Desikan-Killiany atlas to the HCP fsLR-32k_midthickness cortical surface. The named left hemisphere ROIs we analysed were "superior temporal", "banks superior temporal", "supramarginal", "inferior parietal", and "pars opercularis' (see supplementary figures for visualisation).

Semantic models

To assess variation in the correlation of subsets of the data to a model, we used as references semantic models based on three qualitatively different types of information to verify that the pattern of results did not depend on the choice of semantic model. The models consisted of a previously published experiential model of concept representation (Binder et al., 2016), a popular distributional model (word2vec; Mikolov et al., 2013), and a categorical model based on taxonomic category membership. In the experiential model, human ratings on 65 experiential domains are used to represent word meanings in a high-dimensional space. These domains were selected based on known neural processing systems such as colour, shape, visual motion, touch, audition, motor control, olfaction, as well as other fundamental aspects of experience whose neural substrates are less clearly understood, such as space, time, affect, reward, numerosity, and others. The model RDM used in the present study was generated from experiential ratings for the 320 words in Study 1 and 300 words in Study 2. The performance of this model as measured through RSA has been previously reported (Tong et al., 2022). The distributional model uses a neural network trained to predict a masked word using the surrounding words in a context window. We used the pretrained word embeddings accessed through the Python Gensim toolbox (Řehůřek & Sojka, 2010). The categorical model was created to encode the taxonomic structure of the concept sets in both studies. For Study 1 this consisted of two super-categories (events and objects) with 4 sub-categories in each, and for Study 2 the model consisted of 6 categories of concepts (see Figure 1 for visualisation).
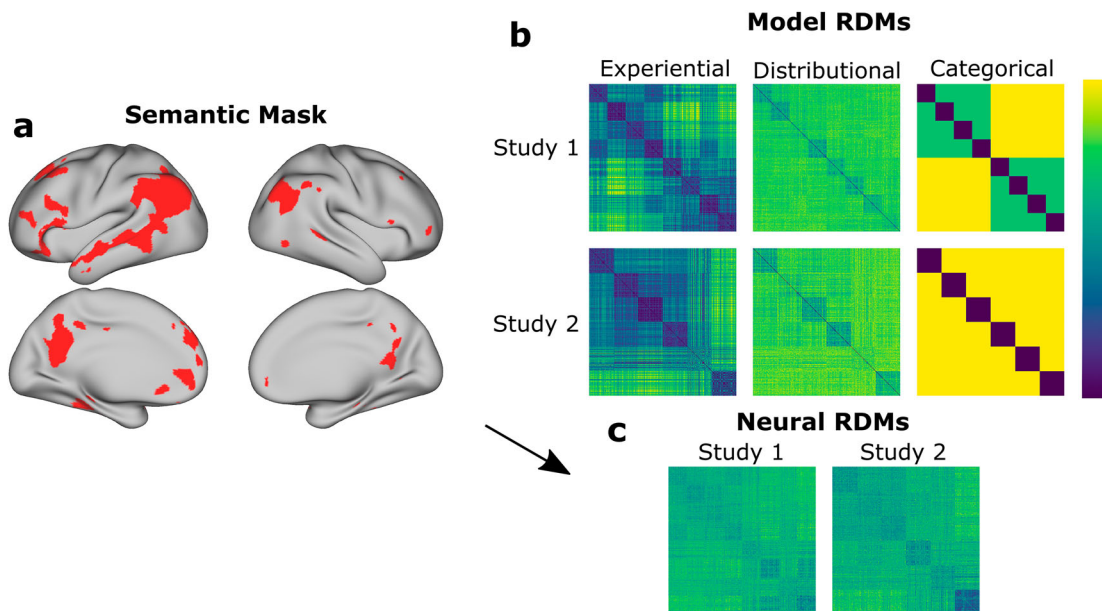

Figure 1. Overview of RSA method. a. The mask used to define the main region of interest in the present study **b.** The model RDMs used in the analyses, with words sorted by semantic category. **c.** Group-averaged neural RDMs from the two data sets analysed.

RSA and Cronbach's alpha

Cronbach's alpha is a reliability coefficient that quantifies the internal consistency of a measure. Values typically range from zero to one, with higher values indicating a more reliable test. Neural RDMs were generated using Pearson correlation distances of vertex-wise z-scored beta-estimates calculated from the regression. As applied to RSA, each participant can be thought of as a test item, and alpha as capturing the overall internal consistency of the group-averaged neural RDM. This fits well with the RSA approach described in the original paper by Kriegeskorte et al. (2008), where single subject neural RDMs are averaged together, and the resulting average RDM is compared to different models. Using group-averaged RDMs when comparing models has the advantage of testing a model's ability to explain only the variance that is common across all participants. Importantly, Cronbach's alpha is also equivalent both to an ICC for the average of measurements across participants, using the two-way random/mixed effect model ANOVA consistency definition (Bravo & Potvin, 1991; McGraw & Wong, 1996), and to the average of all possible split-half tau-equivalent reliabilities for a dataset (Warrens, 2015). It thus provides an index of consistency, or reproducibility, of the results across independent subsets of participants.

Our analysis used the off-diagonal elements of participants' neural RDMs. Following the steps used by the python RSA toolbox when pooling RDMs, we z-scored each participant's RDM values separately prior to calculating alpha (Nili et al., 2014). We used the python Pinguion toolbox (Vallat, 2018) to calculate alpha as a function of the number of stimulus presentations and the number of participants. A random resampling approach was used, in which alpha was calculated 1,000 times for each sample size ranging from 5 to 37 participants, and for 1–6 stimulus presentations at each sample size. We also show the relation between average reliability and average RSA correlation for RDMs derived from a particular combination presentation. Specifically, using RDMs derived from the full set of stimuli presentations, we used 1,000 resamples at each sample size ranging from 5 to 37 participants and calculated the average alpha and correlation to three different semantic models. This relation was then plotted and fitted with a simple square-root function having one degree of freedom for scaling.

We also demonstrated how RDM reliability for a given participant sample size can be estimated based on a smaller sample. To accomplish this, using the method proposed by de Vet and colleagues (2017) to estimate the effect on inter-rater reliability of averaging additional raters, we applied the Spearman-Brown (SB) prophecy formula to the ICC two-way random effect model ANOVA single measurement value calculated from all participants in each study.

Repetition effects

To examine potential repetition suppression effects, the reliability analyses were repeated for each of the six single presentations as well as selected combinations of presentations (e.g. the first presentations of each of the three scanning sessions). In addition, RSA Spearman correlation values between the neural RDMs and the semantic model RDMs were calculated separately for each stimulus set presentation to estimate differences in data quality between the presentations. The RSA correlation was used instead of reliability because the latter is a group-level point estimate, whereas the former allowed us to test for significant differences through a 2-way repeated measures ANOVA using the individual correlation values for each participant. Further, since the correlation between the neural RDM and a model RDM is usually the value of interest in an RSA study, we felt that an understanding of what presentations contained the most signal of interest would be most informative for researchers using RSA. Post-hoc pairwise t-tests were adjusted for multiple comparisons using the Benjamini-Hochberg false-discovery rate (Benjamini & Hochberg, 1995). Correlations with the model were also calculated separately using: 1) only the first presentation from each session (i.e. combination of presentations 1, 3, and 5); 2) only the second presentation from each session; and 3) both presentations from the first session combined with the first presentations from the remaining sessions (i.e. presentations 1, 2, 3, and 5).

Noise ceiling estimates

A common definition of the noise ceiling is the one popularised by the RSA toolbox (Nili et al., 2014). In this formulation, the lower-bound estimate of the noise ceiling is the average correlation between a single participant's RDM and the average RDM from the remaining participants. This metric is an underestimate of the shared variance and can be viewed in contrast to the upper-bound estimate of the noise ceiling, in which the RDM of each participant is correlated with the average RDM of all participants including themselves. This provides an overestimate of the shared variance. As the number of participants increases, both the upper and lower noise ceiling bounds asymptotically approach each other. Intuitively, these values must converge because, as the number of participants

approaches infinity, the contribution of an individual participant to an average tends toward zero. When the noise ceiling is calculated using the Pearson correlation coefficient, and we use the assumptions of classical measurement theory, we can calculate explicit equations governing the expected values of the upper – and lower-bound noise ceiling estimates. We determined the relationship between a point estimate at one sample size and the expected value of the noise ceiling estimate at all other sample sizes. A more detailed motivation for the formula is provided in the Supplemental Text. Given N time series each denoted by Rn, the lower-bound noise ceiling estimate as a function of n subjects follows the form:

$$A = \frac{1}{N} \sum_{n=1}^{N} Var(R_n)$$

$$B = Var\left(\frac{1}{N} \sum_{n=1}^{N} R_n\right)$$

$$C = \frac{A - NB}{1 - N}$$

$$D = \frac{A - B}{1 - \frac{1}{N}}$$

$$LowerNoise(n) = \sqrt{\frac{C}{\sqrt{A} * \sqrt{C + \frac{D}{n - 1}}}}$$

Example code demonstrating how to calculate the expected value of the noise ceiling at different sample sizes can be found in the online GitHub repository associated with this publication. To validate the decomposition of variance implicit in the above

derivation, we plotted the predicted curve against point estimates of the noise ceilings at different sample sizes derived from 1,000 random resamples at sample sizes ranging from 5 to 37.

Results

Cronbach's alpha by number of participants

As shown in Figure 2, in both studies, reliability as measured by Cronbach's alpha increased monotonically with participant sample size. Observed values of alpha closely matched those computed using the Spearman-Brown (SB) Prophecy formula applied to ICC values. Visual inspection of the trend indicates that reliability does not begin to plateau until at least 30 participants are present in the group average. Using all 6 presentations in Studies 1 and 2, a sample size of 20 participants yielded alphas of .58 and .56, and using the complete sample yielded alphas of .74 and .72, respectively. To reach alpha = .8 would require around 58 and 60 participants in Studies 1 and 2 with the current task design. When all 6 presentations were included, for each resample where alpha was calculated we also calculated the RSA correlation values between (group averaged) neural RDMs and three semantic models. Alpha and RSA correlations were averaged at each sample size and plotted against each other demonstrating a square-root relation (Figure 3).

Cronbach's alpha by number of stimulus repetitions

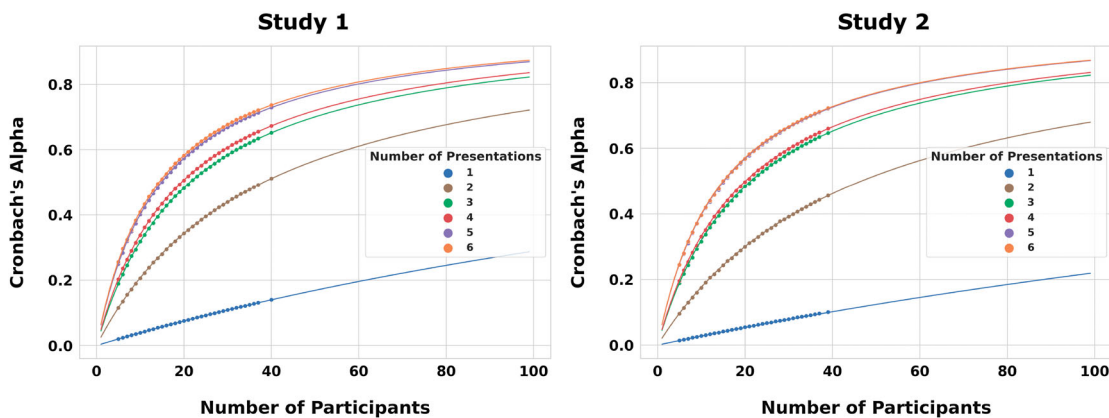Figure 2 also illustrates the degree of improvement in reliability with an increasing number of stimulus

Figure 2. **Reliability as a function of the number of stimulus presentations and the number of participants**. Each point, except for the ones furthest to the right on each plot, represents the average Cronbach's alpha calculated from 1,000 random resamples. The fitted line represents the Spearman-Brown (SB) Prophecy formula applied to the two-way random effects single measurement calculated using neural RDMs from all participants in each presentation combination.
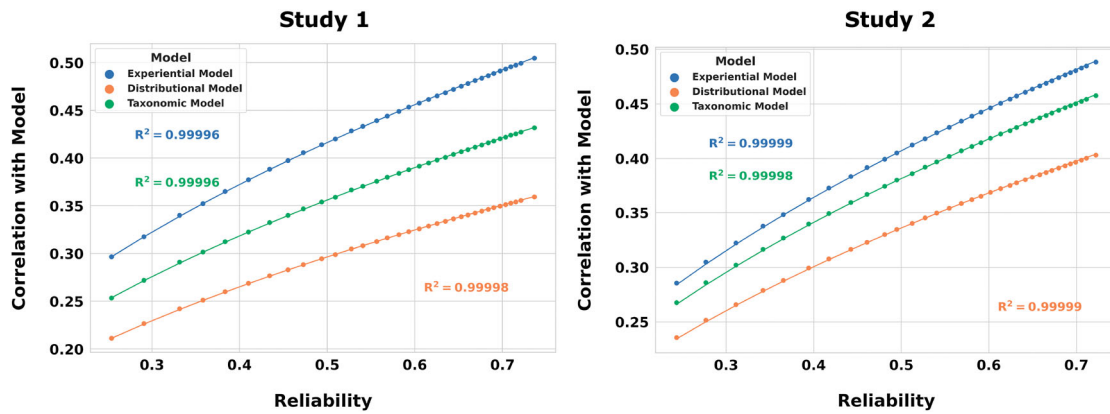
Figure 3. Relation between mean RSA values and Cronbach's alpha. Using RDMs derived from all 6 presentations, each point on the plot, except for the two points with the highest reliability (i.e. furthest to the right), are generated from the average of 1,000 resamples at a particular sample size. The horizontal axis represents Cronbach's alpha, and the vertical axis represents the average RSA correlation between an averaged set of neural RDMs and an RDM derived from a semantic model. The points with the highest reliability are based on the full set of participants and did not involve any resampling. The figure demonstrates that the average RSA correlation is proportional to the square-root of the reliability. This relation, along with the ability to estimate reliability at different sample sizes using the Spearman-Brown formula, allows for estimating group averaged RDM correlations to a model at different sample sizes.

presentations. Alpha remained low when only 1 presentation was included, even with large participant samples. In both studies, the second presentation substantially improved reliability, with an approximately 5-fold increase in alpha values when using two stimulus presentations compared to just one. Smaller but still substantial improvements were produced by the third presentation and the fifth presentation, with hardly any additional gain from the fourth and sixth presentations. This pattern suggests a gradual decline in the added value of repetitions, particularly repetitions occurring in the same session. This overall pattern of findings was also replicated in 5 other ROIs (**Figures S5-S9**).

The between- and within-session repetition effects are seen more clearly in Figure 4, which shows the mean RSA correlation value between the neural RDM and the experiential model RDM for each presentation in each study. In Study 1, both the across-session and within-session main effects were highly significant ($F(2,78) = 31.29$, $p < .001$; and $F(2,39) = 240.82$, $p < .001$, respectively). An interaction effect was also significant ($F(2,78) = 5.56$, $p < .01$). An analysis of the simple effects revealed highly significant within-session differences for all three days, with significantly stronger correlations seen with the first presentation relative to the second (corrected $ps < .001$). Across sessions, the RSA correlation for the first presentation on the first day was significantly stronger than the first presentations on each subsequent day (corrected $ps < .001$). Similarly, the correlation for the second presentation on the first day was stronger than that seen on the subsequent days (corrected $ps < .001$). The declines observed between the second

and third days were smaller in magnitude. The decline was still significant for the second presentation across those days (corrected $p < .05$), but only approached significance for the first presentation ($p < .06$). In Study 2, both within and across session main effects were highly significant ($F(2,76) = 13.90$, $p < .001$; and $F(2,38) = 67.70$, $p < .001$, respectively) with no significant interaction. Follow-up analyses with different models and different ROIs demonstrated the same qualitative effects (**Figure S3** and **Figures S5-S9**, respectively). We also analysed the participant response times and found main effects of stimulus repetition for both within- and across-session repetition in Study 1 (F (2,78) = 110.0, $p < .001$; and F(2,39) = 3.58, $p < .05$) and a main effect of within-session repetition in Study 2 (F (2,38) = 130.7, $p < .001$; **Figure S2**), which mirrored the repetition suppression patterns observed in the fMRI data.

These effects are illustrated in a different way in Figure 5, which shows mean RSA neural-model correlation values for different combinations of presentations. The large effect of within-session repetition is illustrated by the much greater correlations using presentations 1, 3, and 5 (first presentation in each session) compared to presentations 2, 4, and 6. The correlation using the first presentations in each session (1, 3, 5) was slightly but significantly higher than the correlation obtained from presentations 1–3 in Study 1 (mean difference = .010, $p < .005$), with no statistically significant difference in Study 2. Leaving out data from later within-session repetitions (i.e. presentations 4 and 6) resulted in correlation values (r = 0.138 and r = .130) that were nearly as high as
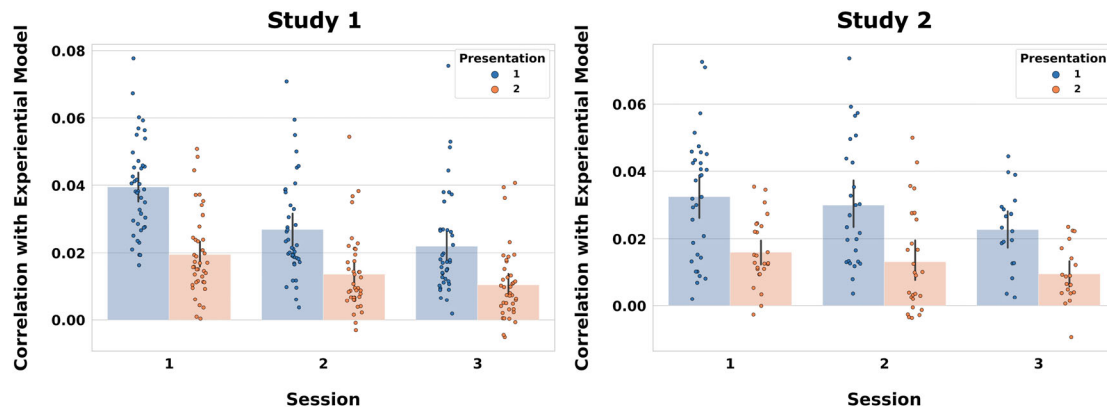
Figure 4. **Spearman correlation with experiential model for each stimulus presentation.** Each point represents the correlation between a participant's neural RDM and an RDM derived from the experiential semantic model. The bars show the average RSA values, and the error bars show the 95% confidence intervals based on 1,000 bootstraps. The across-session and within-session main effects were highly significant, and an interaction effect was also significant in study 1.

with all 6 presentations (r = 0.15 and r = 0.136), though this small difference was statistically reliable in both studies (mean Study 1 difference = .01, $p < .00001$; mean Study 2 difference = .006, $p < .05$). Similar patterns of results for presentation combinations, with some small variation in relative order of combinations (1-2-3-4) and (1-3-5), were found in the analysis of 5 other ROIs (**Figures S5-S9**). The overall pattern of results seen in Figure 5 were also observed when using reliability point estimates instead of RSA correlations (**Figure S4**).

Noise ceiling by number of participants

Analysis of the upper and lower bound estimates of the noise ceiling showed asymptotic behaviour (Figure 6). For Study 1, with 10 participants included in the analysis, the mean lower noise ceiling had a value of 0.16; with 20

participants, the value increased to 0.19; and with all participants, the value was 0.22. For study 2, with 10 participants included in the analysis, the mean lower noise ceiling had a value of 0.15; with 20 participants, the value increased to 0.18; and with all participants, the value was 0.21. With all participants included in the analyses, the differences between the upper and lower noise ceilings were .08 in both studies. The resampled lower noise ceiling was well fit by our model equation. Extrapolation of these curves shows that they converge to 95% of their asymptotic value at sample sizes of 134 and 140 for Studies 1 and 2, respectively.

Discussion

Our findings demonstrate significant within- and across-session repetition suppression effects on the RSA correlation between neural and model RDMs. Thus, these
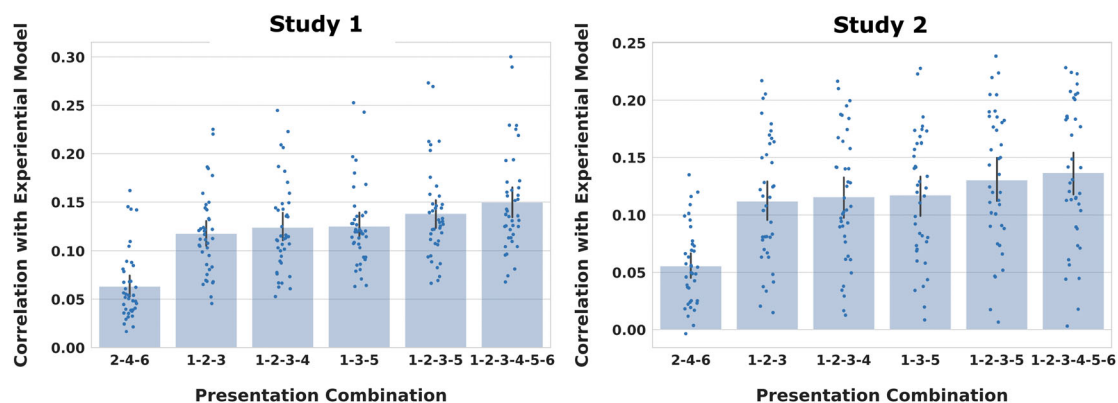

Figure 5. **Neural RDM correlations as a function of selected presentation combinations.** Each point represents the Spearman correlation between a single participant's neural RDM and the experiential model. The bars show the average RSA values, and the error bars show the 95% confidence intervals based on 1,000 bootstraps. The neural RDMs were derived using only the functional scans for the presentations noted on the horizontal axis. The data indicate that excluding the second presentations in sessions two and three only modestly affects the observed RSA correlation.
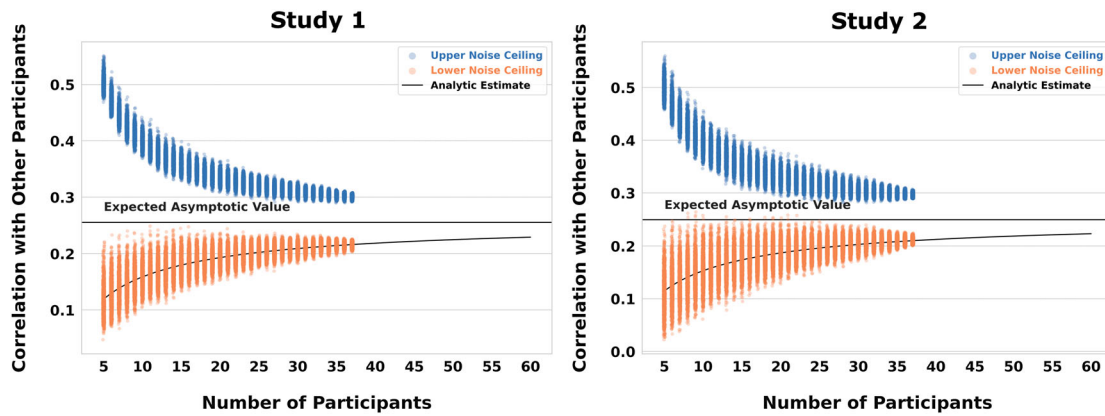
Figure 6. Upper- and lower-bound noise ceiling estimates by number of participants. Each point in this plot represents a noise ceiling estimate for a randomly sampled group of participants, with 1,000 estimates for each sample size. Also plotted is the expected asymptotic value of the upper- and lower-bound noise ceilings, as well as an analytic estimate of the lower-bound noise ceiling as a function of the number of participants extended out to a sample size of 60 participants. The figure shows a relative plateauing of the noise ceiling estimates at around N = 25 in both studies.

results highlight the importance of another design consideration beyond simply the number of repetitions, namely, whether the stimulus repetitions occur within the same fMRI scanning session or across separate days. While increasing the number of stimulus presentations generally improved the reliability of the group-averaged RDM, there were substantial adaptation, or repetition suppression, effects, which were most pronounced for within-session repetitions, with some release from adaptation observed across days. Additional presentations of the stimulus within the same fMRI scanning session appear to produce minimal benefit after the first day.

The magnitude and duration of the repetition suppression effect are notable. Repetition suppression is rarely mentioned as a potential problem in RSA studies that employ stimulus repetition, and it has likely been underestimated in MVPA studies more broadly. For example, Zhang et al. (2022) stated that "repetition suppression effects in fMRI are short-lived, dissipating on the order of seconds, and are strongest when few other stimuli are presented between repetitions [...]." Their study used 32 stimuli presented 6 times on the same day; we found a large within-session repetition suppression effect even with 320 stimuli presented only twice on each day. Further, diminishing returns across scanning sessions were notable even with an average interval between the first and last session of approximately 45 days (**Figure S1**). Similar repetition suppression effects were seen across two separate data sets using different stimuli and participants as well as across multiple ROIs. Importantly, such effects are likely to be much larger in studies that use smaller stimulus sets or more than two repetitions in a single day.

In our condition-rich, item-level RSA design, we found limits and tradeoffs regarding sample size and number of stimulus presentations analogous to those previously reported in univariate fMRI (Nee, 2019). Specifically, very large sample sizes would be required to achieve adequate reliability with fewer than three stimulus presentations. In fact, increasing sample size was of minimal benefit with only one presentation, as reliability remained low even with sample size projections over 100. Stimulus repetition, therefore, is a useful strategy for enhancing reliability and can be "traded" for a smaller sample size. In our data, for example, the alpha obtained with 40 participants and 2 presentations (∼.5) could have been reached with half as many participants and a third presentation on a different day. However, beyond the third presentation, there were diminishing returns to adding more repetitions, particularly within a session. While the highest alphas and RSA correlations were seen with the full 6 presentations, the results suggest that for these two data sets, it would have been possible to obtain nearly indistinguishable results with only 4 of the 6 presentations by eliminating a second within-session presentation of the stimuli after the first session, thus substantially reducing total scanning time.

To be clear, the sample size curves presented here are not intended to provide hard recommendations for sample size. The absolute values of alpha obtained in any given study likely depend on several factors, including the number and type of stimuli used, the task performed, and the image analysis methods. The overall pattern of sample size and repetition effects, however, was observed to generalise across two studies using different stimuli and participants, and across ROIs

differing in size and location, thus we believe that these patterns are likely to generalise to other studies using an item-level RSA approach. As mentioned earlier, the repetition suppression effects we observed are likely to be even larger in studies using smaller stimulus sets and more closely spaced repetitions. Furthermore, it seems likely that the advantages of having at least 3 stimulus presentations and the general pattern of trade-offs between sample size and stimulus presentations would apply even to studies using other tasks and stimulus types.

Because it is not possible to make universal recommendations for the amount of individual-level data and sample size that would be appropriate to all experimental designs, Nee (2019) suggested that researchers examine these factors in their own pilot data and estimate the optimal amount of data and sample size from those data. The current findings provide support for the application of ICCs and the Spearman-Brown prophecy formula to the group-averaged neural RDM at intermediate points in data collection in order to estimate the potential gains in reliability to be expected by increasing sample size given the obtained level of interindividual consistency. This usage was proposed by de Vet et al. (2017) as a method to estimate the number of raters needed to attain an adequate level of interrater reliability. The current findings extend that usage to estimates of sample size for studies using neural RDMs. We also derived a formula relating how the upper- and lower-bound estimates of the noise ceiling change as a function of the number of participants.

We chose to assess the reliability of the group-averaged neural RDM because it is a function of both between-participant reliability and the number of participants. While individual participant neural RDMs are typically used for statistical inference in RSA, the power to detect a significant correlation depends on both the reliability of the individual RDMs, captured in the ICC two-way random effects, single measurement value, and the number of participants, captured in the Spearman-Brown prophecy formula applied to that ICC value. Beyond this, the correlation between a model RDM and a group averaged neural RDM is more closely related to how well a model accounts for "shared" or "explainable" variance.

The calculation of alpha or ICCs may provide additional insights regarding reliability and can be interpreted within a broader context. However, we note that the upper and lower noise ceilings have the benefit of being applicable to many distance metrics (e.g. Pearson, Spearman, Kendal-tau-a, Jaccard), whereas classical reliability measures are often not applicable when considering measures other than Pearson correlations.

There was relatively slow convergence of upper and lower noise ceilings within the range of sample sizes typically used in fMRI. We do not believe that the persistence of a sizable difference between the upper and lower bound estimates of the noise ceiling is particular to our data, and it is likely the case that additional participants would only marginally reduce this difference at common sample sizes. Although not explicitly reported in the results (because resampling variability is difficult to assess when subsampling from a finite pool of participants), we note that that the variability in the lower noise ceiling estimate was high when fewer than 10 participants were used in its calculation. However, this large variability is also compounded by the fact that the non-independence of subsets means that the variance observed at any given sample size is an underestimate of the variance that would be found if a larger dataset was used.

One tacit assumption throughout this paper has been that the observed changes in reliability are driven by the information of interest (in our case, semantic information), and that higher reliability is desirable. However, as sometimes occurs in fMRI, very reliable effects could simply reflect nuisance factors that are consistent across participants, such as the blocking structure of trials (Cai et al., 2019). While there are almost certainly sources of shared variance in our study that are not related to the (semantic) information of interest, the similarity between the repetition suppression patterns found in the reliability curves and in the analyses using semantic models indicates that the observed changes in reliability likely reflect changes in the amount of semantic information. This is likely to be the case only when nuisance variables (such as the blocking structure of trials) are reasonably well controlled.

In closing, while much work remains to be done to thoroughly characterise the factors that affect the reliability of neural RDMs, the present analyses provide some initial steps in this process, and we hope they will encourage the reporting of intermediate reliability statistics to aid in comparisons across studies. In the case of RSA, this amounts to reporting a standard measure of reliability for neural RDMs along with the correlation measured.

Acknowledgements

Data availability statement

The neural RDMs as well as code to generate all figures are available at the following GitHub repository: https://github.com/smazurchuk/rsa_reliability

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

ORCID

Stephen Mazurchuk http://orcid.org/0000-0002-9057-6382

References

Allefeld, C., & Haynes, J.-D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage*, *89*, 345–357. https://doi.org/10.1016/j.neuroimage.2013.11.043. PMID - 24296330.

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. S. (2016). Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, *128*, 44–53. https://doi.org/10.1016/j.neuroimage.2015.12.035. PMID - 26732404.

Arbuckle, S. A., Yokoi, A., Pruszynski, J. A., & Diedrichsen, J. (2019). Stability of representational geometry across a wide range of fMRI activity levels. *NeuroImage*, *186*, 155–163. https://doi.org/10.1016/j.neuroimage.2018.11.002. PMID - 30395930.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3-4), 130–174. https://doi.org/10.1080/02643294.2016.1147426. PMID - 27310469.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055. PMID - 19329570.

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2020). Measures of neural similarity. *Computational Brain & Behavior*, *3*(4), 369–383. https://doi.org/10.1007/s42113-019-00068-5. PMID - 33225218.

Borghesani, V., Pedregosa, F., Buiatti, M., Amadon, A., Eger, E., & Piazza, M. (2016). Word meaning in the ventral visual path: A perceptual to conceptual gradient of semantic coding. *NeuroImage*, *143*, 128–140. https://doi.org/10.1016/j.neuroimage.2016.08.068. PMID - 27592809.

Bravo, G., & Potvin, L. (1991). Estimating the reliability of continuous measures with cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. *Journal of Clinical Epidemiology*, *44*(4-5), 381–390. https://doi.org/10.1016/0895-4356(91)90076-L. PMID - 2010781.

Bruffaerts, R., Dupont, P., Peeters, R., Deyne, S. D., Storms, G., & Vandenberghe, R. (2013). Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *The Journal of Neuroscience*, *33*(47), 18597–18607. https://doi.org/10.1523/JNEUROSCI.1548-13.2013. PMID - 24259581.

Cai, M. B., Schuck, N. W., Pillow, J. W., & Niv, Y. (2019). Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLOS Computational Biology*, *15*(5), e1006299. https://doi.org/10.1371/journal.pcbi.1006299. PMID - 31125335.

Carota, F., Nili, H., Pulvermüller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: Evidence from RSA of BOLD signals. *NeuroImage*, *224*, 117408. https://doi.org/10.1016/j.neuroimage.2020.117408

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014. PMID - 8812068.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021. PMID - 16530430.

de Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman–brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45–49. https://doi.org/10.1016/j.jclinepi.2017.01.013. PMID - 28342902.

Dobbins, I. G., Schnyer, D. M., Verfaellie, M., & Schacter, D. L. (2004). Cortical activity reductions during repetition priming can result from rapid response learning. *Nature*, *428*(6980), 316–319. https://doi.org/10.1038/nature02400. PMID - 14990968.

Dong, J., Li, A., Chen, C., Qu, J., Jiang, N., Sun, Y., Hu, L., & Mei, L. (2021). Language distance in orthographic transparency affects cross-language pattern similarity between native and non-native languages. *Human Brain Mapping*, *42*(4), 893–907. https://doi.org/10.1002/hbm.25266. PMID - 33112483.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, I. A., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4. PMID - 30532080.

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6),

e2108091119. https://doi.org/10.1073/pnas.2108091119. PMID - 35115397.

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S. P., & Tamez, E. R. (2017). Decoding levels of representation in reading: A representational similarity approach. *Cortex*, *90*, 88–102. https://doi.org/10.1016/j.cortex.2017.02.017. PMID - 28384482.

Gao, Z., Zheng, L., Gouws, A., Krieger-Redwood, K., Wang, X., Varga, D., Smallwood, J., & Jefferies, E. (2022). Context free and context-dependent conceptual representation in the brain. *Cerebral Cortex*, https://doi.org/10.1093/cercor/bhac058. PMID - 152 - 166.

Glasser, Matthew F, Sotiropoulos, Stamatios N., Wilson, J. Anthony, Coalson, Timothy S, Fischl, Bruce, Andersson, Jesper L., Xu, Junqian, Jbabdi, Saad, Webster, Matthew, Polimeni, Jonathan R, Van Essen, David C., & Jenkinson, Mark. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006. PMID - 16321563.

Guo, W., Geng, S., Cao, M., & Feng, J. (2022). Functional gradient of the fusiform cortex for Chinese character recognition. *eNeuro*, *9*(3), ENEURO.0495–21.2022-0421.2022. https://doi.org/10.1523/eneuro.0495-21.2022. PMID - 35545424.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), https://doi.org/10.3389/neuro.06.004.2008. PMID - 19104670.

Lee, S.-M., Henson, R. N., & Lin, C.-Y. (2020). Neural correlates of repetition priming: A coordinate-based meta-analysis of fMRI studies. *Frontiers in Human Neuroscience*, *14*, 565114. https://doi.org/10.3389/fnhum.2020.565114. PMID - 33192395.

Li, H., Cao, Y., Chen, C., Liu, X., Zhang, S., & Mei, L. (2023). The depth of semantic processing modulates cross-language pattern similarity in Chinese–English bilinguals. *Human Brain Mapping*, 2085. https://doi.org/10.1002/hbm.26195

Li, H., Qu, J., Chen, C., Chen, Y., Xue, G., Zhang, L., Lu, C., & Mei, L. (2019). Lexical learning in a new language leads to neural pattern similarity with word reading in native language. *Human Brain Mapping*, *40*(1), 98–109. https://doi.org/10.1002/hbm.24357. PMID - 30136328.

Liu, X., Hu, L., Qu, J., Zhang, S., Su, X., Li, A., & Mei, L. (2023). Neural similarities and differences between native and second languages in the bilateral fusiform cortex in Chinese-English bilinguals. *Neuropsychologia*, *179*, 108464. https://doi.org/10.1016/j.neuropsychologia.2022.108464. PMID - 36565993.

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

Meersmans, K., Storms, G., Deyne, S. D., Bruffaerts, R., Dupont, P., & Vandenberghe, R. (2022). Orienting to different dimensions of word meaning alters the representation of word meaning in early processing regions. *Cerebral Cortex*, bhab3302-. https://doi.org/10.1093/cercor/bhab416. PMID - 34963135.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1301.3781.

Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, *2*(1), 130. https://doi.org/10.1038/s42003-019-0378-6. PMID - 30993214.

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. https://doi.org/10.1371/journal.pcbi.1003553. PMID - 24743308.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4. PMID - 5146491.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y. PMID - 30734206.

Řehůřek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora*.

Ritchie, J. B., Masson, H. L., Bracci, S., & Beeck, H. P. O. D. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage*, *245*, 118686. https://doi.org/10.1016/j.neuroimage.2021.118686. PMID - 34728244.

Staples, R., & Graves, W. W. (2020). Neural components of reading revealed by distributed and symbolic computational models. *Neurobiology of Language*, *1*(4), 381–401. https://doi.org/10.1162/nol_a_00018. PMID - 36339637.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, *35*(1), 105–120. https://doi.org/10.1016/j.neuroimage.2006.11.054. PMID - 17239619.

Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A distributed network for multimodal experiential representation of concepts. *The Journal of Neuroscience*, *42*(37), 7121–7130. https://doi.org/10.1523/JNEUROSCI.1243-21.2022. PMID - 35940877.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), 62. https://doi.org/10.1038/s42003-018-0073-z. PMID - 30271944.

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012. PMID - 26707889.

Wang, X., Xu, Y., Wang, Y., Zeng, Y., Zhang, J., Ling, Z., & Bi, Y. (2018). Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Scientific Reports*, *8*(1), 3047. https://doi.org/10.1038/s41598-018-21062-0. PMID - 29445098.

Warrens, M. J. (2015). Springer proceedings in mathematics & statistics. *Springer Proceedings in Mathematics & Statistics*, 293–300. https://doi.org/10.1007/978-3-319-07503-7_18

Zhang, Y., Lemarchand, R., Asyraff, A., & Hoffman, P. (2022). Representation of motion concepts in occipitotemporal cortex: fMRI activation, decoding and connectivity analyses. *NeuroImage*, *259*, 119450. https://doi.org/10.1016/j.neuroimage.2022.119450. PMID - 35798252.