SUPPLEMENTAL DATA

Parametric coding of the size and clutter of natural scenes in the human brain

Soojin Park, Talia Konkle, & Aude Oliva

Table of contents:

Supplementary Analyses:

Supplementary Methods:

Supplementary Analysis: SVM classification and confusion matrix for all ROIs

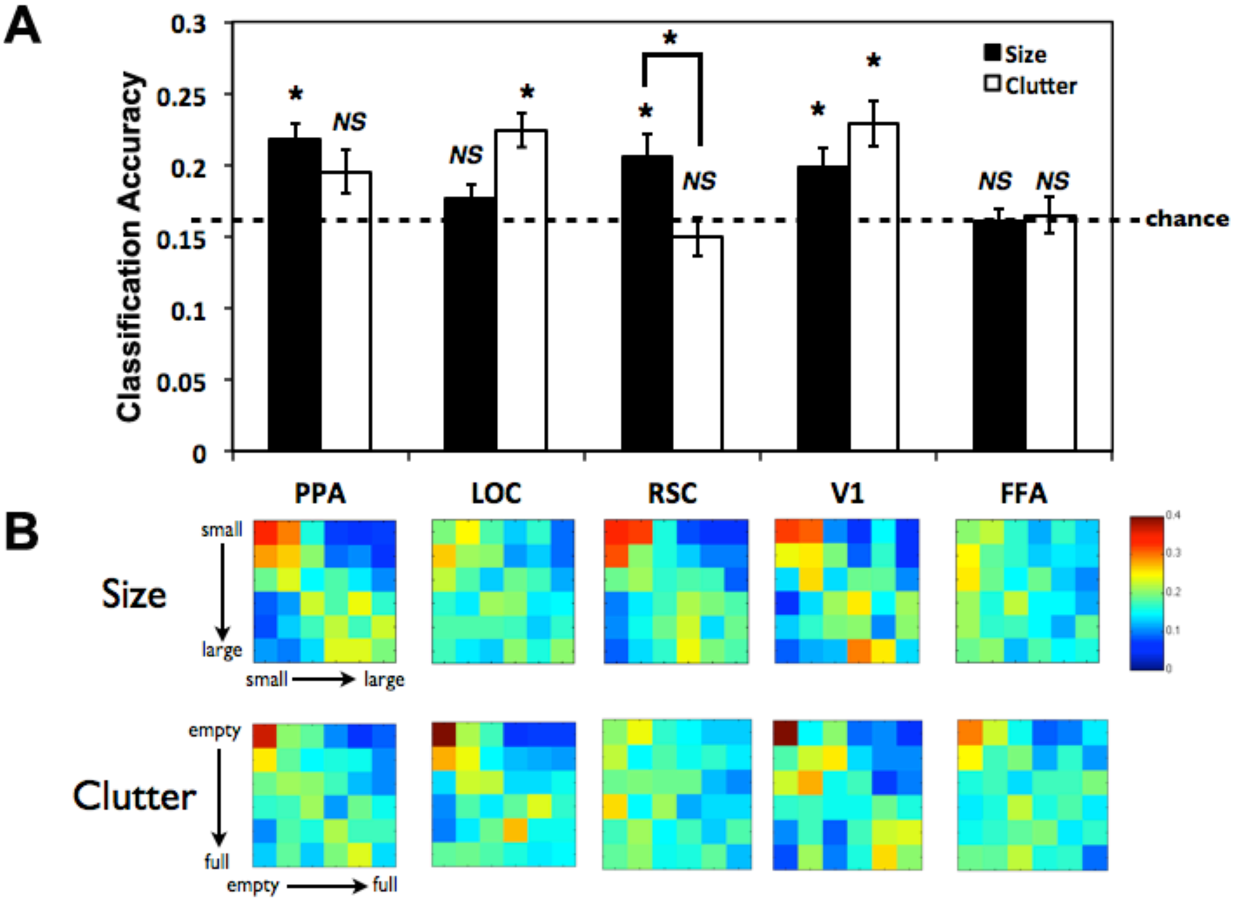

Figure S1. A. Standard SVM classification accuracy for the size and clutter levels. In addition to the ridge regression analysis, we also performed a standard multivariate pattern classification analysis using linear support vector machine (SVM) classifier using LIBSVM. For each block, we computed the average pattern of activity across the voxels. We used a "leave-one-entire-category-out" cross validation in which we used different sets of categories representing the same size of a scene as the training and test samples. For example, an SVM classifier was trained to classify the six size levels using all categories but one, and then was tested on the remaining category. This was repeated across all samples so that each sample of the dataset played a role in training and testing. The procedure was repeated for the clutter dimension as well. Percent correct classification for each subject and each ROI was calculated as the average performance over the cross-validation iterations.

When the classifier was tested for the size of a scene, the PPA, RSC, and V1 could classify the size of a novel scene category above chance (PPA: t(11) = 4.47, *p* < .01; RSC: t(10) = 2.49, *p* < .05), while LOC could not (t(9) = 1.13, *p* > .28). Such accurate classification requires the generalization of representation across semantic categories, since the classifier was always tested on a novel semantic category. On the contrary, when the classifier was tested for the amount of clutter with the same exact sets of scenes, LOC performed significantly above chance (t(9) = 4.84, *p* <.01), while the PPA only marginally classified the clutter level (t(11) = 1.85, *p* =

.09) and RSC could not classify clutter levels at all (t(10) = -1.271, *p* = .23). V1 could classify the clutter levels correctly (t(11) = 3.88, *p* < .01). Interestingly, RSC showed specificity for the representation of scene size and not clutter (paired t-test t(10) = 2.89, *p* < .02), while LOC showed a specificity for the representation of clutter and not size (paired t-test t(9) = -2.83, *p* < .05). There was a significant interaction across the RSC and LOC for the size and clutter (F(1,8)=17.49, *p* < .01).

B. Confusion matrices of the size and clutter classifier. The confusion matrix in the PPA and RSC for size shows a ‘fuzzy diagonal’, which suggests that confusion errors in the classifier were not random, and instead reflect a sensitivity to scene size (e.g., the classifier made more errors between adjacent levels, creating a fuzzy diagonal in the confusion matrix).

Supplementary Analysis: Parametric pattern analysis overall accuracy results

ROI	Size		Clutter		Size vs Clutter
	%Correct (sem)	difference from chance (chance = 16.7%)	%Correct (sem)	difference from chance (chance = 16.7%)	
PPA	25.2% (1.2)	**t(11)=7.4, p<0.001***	22.5% (1.2)	**t(11)=4.3, p<0.005***	**t(11)=1.9, p=0.09 ~**
LOC	20.1% (1.1)	**t(9)=3.1, p<0.05***	20.9% (1.1)	**t(9)=5.8, p<0.001***	t(9)=-0.8, p=0.46
RSC	26.3% (0.7)	**t(10)=14.4, p<0.001***	17.9% (0.7)	t(10)=1.1, p=0.28	**t(10)=8.9, p<0.001***
V1	23.1% (1.0)	**t(10)=6.3, p<0.001***	24.3% (1.0)	**t(10)=5.4, p<0.001***	t(10)=-0.6, p=0.53
FFA	17.8% (1.0)	t(11)=1.1, p=0.29	19.6% (1.0)	**t(11)=2.5, p<0.05***	t(11)=-1.1, p=0.29

Table S1. Parametric pattern analysis overall accuracy results. In addition to the correlation between the predicted size/ clutter levels and the actual size/clutter levels, we also computed overall percent correct, in which we rounded the predicted level and compared it to the actual level. Chance in this method is 1/6 or 16.7%. The overall accuracy and standard error of the mean for each ROI is shown in Table S1. The table also shows the statistical tests between each ROI and chance, as well as the t-tests between performance on size versus clutter dimension.

Overall percent accuracies were on the range of 17.8% to 26.3%. We observed that all areas performed above chance classification for the size dimension except for the FFA; and all regions performed above chance classification for the clutter dimension except RSC. The RSC region was the only region to show significantly different classification performance between size and clutter, with a trend for this in PPA.

Supplementary Analysis: Permutation Analysis

To measure the probability of obtaining high correlations between predicted and actual levels of size or clutter by chance, we conducted a permutation analysis. For each subject, for a given ROI, we shuffled the scene labels and computed the average r-value following the 6-fold validation procedure. We repeated this procedure 100x to obtain a distribution of r-values. A p-value was obtained by fisher-z transforming the distribution of r-values obtained from 100 shuffled iterations along with the actual r-value for size (or clutter), and computing the z-score of the actual size or clutter regression coefficient with respect to the shuffled distribution. This procedure assumes a normal distribution of transformed r-values, which was appropriate given the shuffled distributions we observed; a non-parametric statistical procedure would require more samples than was computationally feasible. This analysis was performed for the PPA, RSC, and the FFA.

Comparing the r-value obtained when the labels are not shuffled allows us to estimate whether the regression pattern analysis was significantly different from a shuffled baseline in each ROI of individual participant. We found that, in the PPA, 11 of 12 participants showed a significant correlation for size and 9 of 12 participants showed a significant correlation for clutter. In the RSC, 11 of 11 participants showed a significant correlation for size and 2 of 12 participants showed a significant correlation for clutter. Interestingly, in the FFA, 2 of 12 showed a significant correlation for size and 3 of 12 for clutter. Importantly, across all cases across subjects and ROIs, the mean of the shuffled distribution was centered on zero. Together, these results indicate that (i) a correlation of zero is expected by chance, and (ii) at the single subject level, the regression model performance largely follows the results found presented at the group level.

Supplementary Analysis: Mean-Centered Parametric Pattern Analysis

To eliminate the influence of main effects on multi-voxel analysis, we mean-centered each pattern for each scene category in each ROI (so that the average beta was zero across the voxels in the region for each scene category pattern). When we re-analyzed these zero-mean centered patterns in ridge regression analysis, we found the same overall pattern of results. That is, the patterns of the PPA were able to predict both the size and clutter properties, but were significantly better at predicting the size of the scene ($t(11) = 2.76$, $p<.05$); the patterns of the RSC were much more sensitive to the size than clutter ($t(10)=6.48$, $p<.001$). There was a significant interaction across the two ROIs, where RSC showed a significantly larger size vs. clutter difference than the PPA ($F(1,43) = 11.3$, $p=.007$). These direct replications of the results using mean-centered multi-voxel patterns show that results that we obtained in the main manuscript can not be reduced to the main effect.

Comparing the original and mean-centered results directly revealed that the original ROI data showed very slightly but reliably higher correlations for size, but not clutter, dimensions. Size paired t-test: PPA vs PPA-MeanCentered: $t(11)=2.44$, $p=0.033$ *; Clutter paired t-test: PPA vs PPA-MeanCentered: $t(11)=1.99$, $p=0.072$ (n.s.); Size paired t-test: RSC vs RSC-MeanCentered $t(10)=3.15$, $p=0.010$ *; Clutter paired t-test: RSC vs RSC-MeanCentered: $t(10)=-1.21$, $p=0.255$ (n.s.). These likely reflect the contribution of the main effects of the ROI overall response to size.

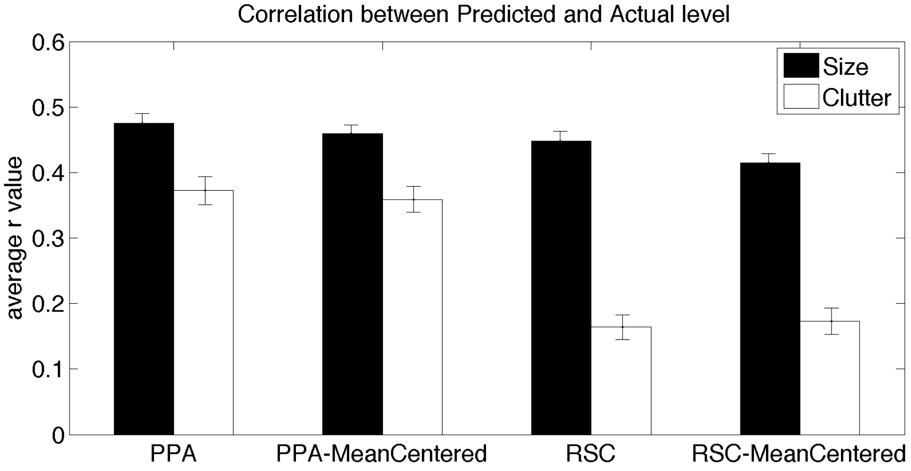

Figure S2. Original and mean-centered multi-voxel regression results in the RSC and PPA. Regions are along the x-axis and the average correlation for size and clutter dimensions across participants is plotted on the y-axis. Error bars reflect +/- 1 within-subject standard error of the mean.

Supplementary Analysis: Anterior and posterior PPA pattern analysis

We examined the degree of size and clutter information present in the multi-voxel patterns of the anterior and posterior subdivisions of the PPA. The anterior and posterior aspect of the PPA showed similar correlations with size and clutter levels as the full ROI (Figure S3). Overall, there was no main effect of anterior/posterior regions (F(1,47)=0.5, p=0.503), nor was there a main effect of size vs clutter (F(1,47)=3.0, p=0.112). However both subregions show a trend for higher correlations for the size dimension. Further, there was no interaction between anterior/posterior and size/cluster dimension (F(1,47)=1.1, p=0.325).

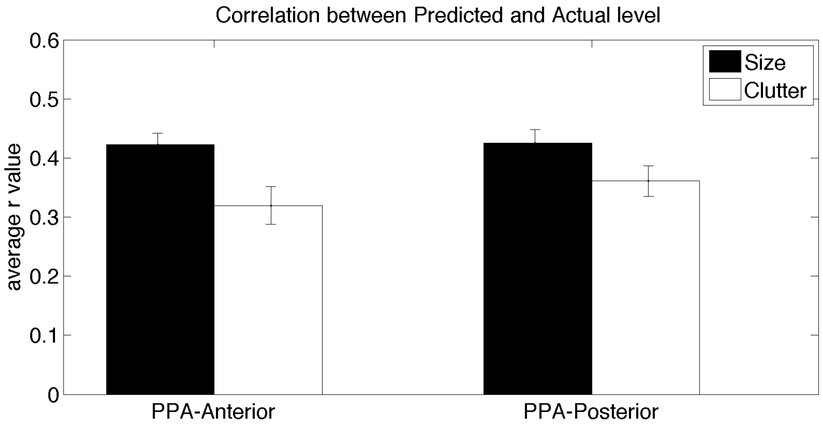

Figure S3. Split-ROI parametric pattern analysis. The PPA was divided into an anterior and posterior section (x-axis) and the average correlation for size and clutter dimensions across participants is plotted on the y-axis. Error bars reflect +/- 1 within-subject standard error of the mean.

Supplementary Methods: Stimulus set validation experiment

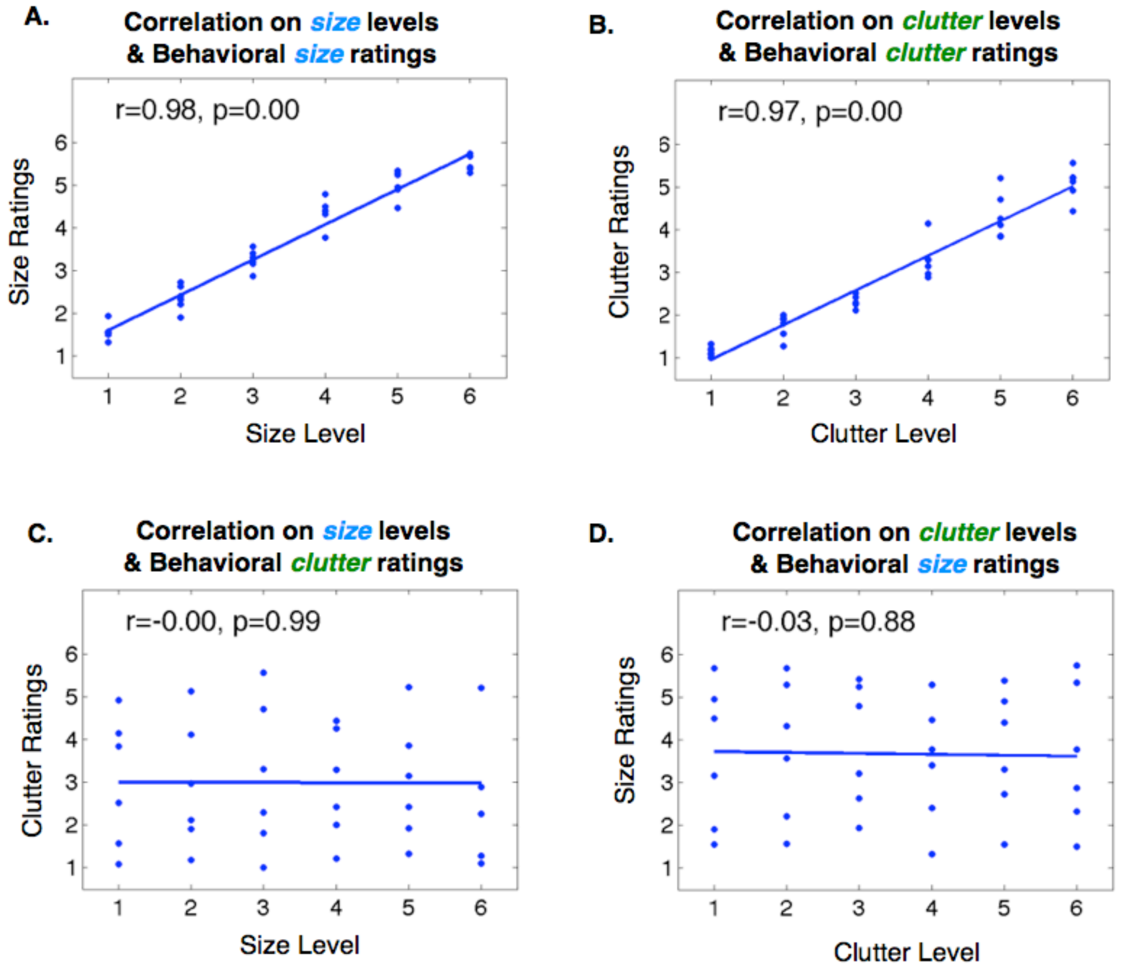

Figure S4. Correlation on pre-selected size and clutter levels and behavioral ratings. A. Correlation between size levels and size behavioral ratings. **B.** Correlation between clutter levels and clutter behavioral ratings. **C & D.** Comparison between size levels and clutter behavioral ratings, and vice versa

In order to validate the size and clutter levels of the stimulus set used in Experiment 2, we obtained 20 ratings for the 432 images (36 categories * 12 exemplars) on both the size and clutter dimensions. Participants were recruited through Amazon Mechanical Turk. Each trial, an image was shown on the screen. In the size experiment, the instructions were to judge the size of the space on a scale of 1 to 6, where 1=very small and 6=very big. In the clutter experiment, the instructions were to judge the clutter in the scene on a scale from 1-6, where 1=empty and 6=fully cluttered. In total, we collected 20 judgments for both size and clutter dimensions per each of 432 individual images.

We compared the pre-selected size and clutter values from the stimulus set (“levels”) with size and clutter average ratings from the behavioral experiments (“ratings”). We observed that there was a strong correlation between the pre-selected levels and the behavioral ratings at the categorical level, for both size (r=.98, p<0.01) and clutter (r=0.97, p<0.01) dimensions (Figure S4 A & B). Consistent with the orthogonal nature of the design, there was no relationship between the size levels and clutter ratings or vice versa (r=0.0, r=-0.3; Figure S4 C & D), and the size ratings and the clutter ratings were not correlated with each other (r=-0.3, all p's > .2). Finally, we found that the ratings were quite consistent across the images within each category: the item standard deviation was 0.30 for size and 0.32 for clutter, averaged over the 36 scene categories. In other words, across the items within each category, the standard deviation of the rankings was less that half of a level on the 6-point scale.

Supplementary Methods: Simulation analysis for ridge regression

We performed a simulation to verify that our hold-out method in the ridge regression train/test procedure leads to unbiased estimates of true size and clutter relationships. We considered how different leave-out methods were able to recover underlying structure, considering several possible relationships between size and clutter. Here we show the results for 3 such data sets: when the data contained a perfect parametric modulation of one dimension and no modulation of the other, vice versa, or perfect modulations of both dimensions. We simplified the dimensionality of the data, considering a 1-voxel ROI. The different leave-out methods were:

(1) *random* item per level: for each size level, leave out one of the 6 categories randomly, ignoring clutter levels

(2) *latin* square leave-out: for each size level, leave out one of the 6 categories so that, across all 6 levels of size, one of each of the 6 levels of clutter has been left out.

(3) *diagonal* leave out: for each size level, leave out the scene with the corresponding level of clutter (e.g. for size level 1, leave out category with clutter level 1, etc.)

(4) *opposite diagonal* leave out: for each size level, leave out the scene with the "opposite" level of clutter (e.g. for size level 1, leave out category with clutter level 6, etc.)

(5) *orthogonal*-leave-out: for each size level, leave out a specific level of clutter and repeat this six times. (e.g. on iteration 1 leave out the categories with clutter level 1; on iteration 2, leave out the categories with clutter level 2, etc.)

For each data set and each leave-out method, we (i) simulated a noisy voxel based on the true parametric relationship across size and clutter dimensions, (ii) fit a regression model (a weight on the voxel) to predict the training levels given the voxel's response; (iii) computed the predicted levels of the held-out test set based on this model, (iv) correlated the predicted levels with the actual levels, (v) and summarized the results as the average fisher-z transformed correlation over 1000 iterations.

Figure S5 summarizes the results. We found that only the *orthogonal*-leave-out procedure returned unbiased estimates of the size and clutter dimensions. Random category leave out methods correctly estimated the informative dimension, but showed systematic negative correlations for the non-informative dimension (rather than zero correlations, as would be expected from the structure of the data). The same was true for a latin-square leave out procedure and a diagonal leave-out procedure. These simulation results hold over a range of noise levels in the voxel's response. Based on these simulations, we used the *orthogonal*-leave-out procedure.

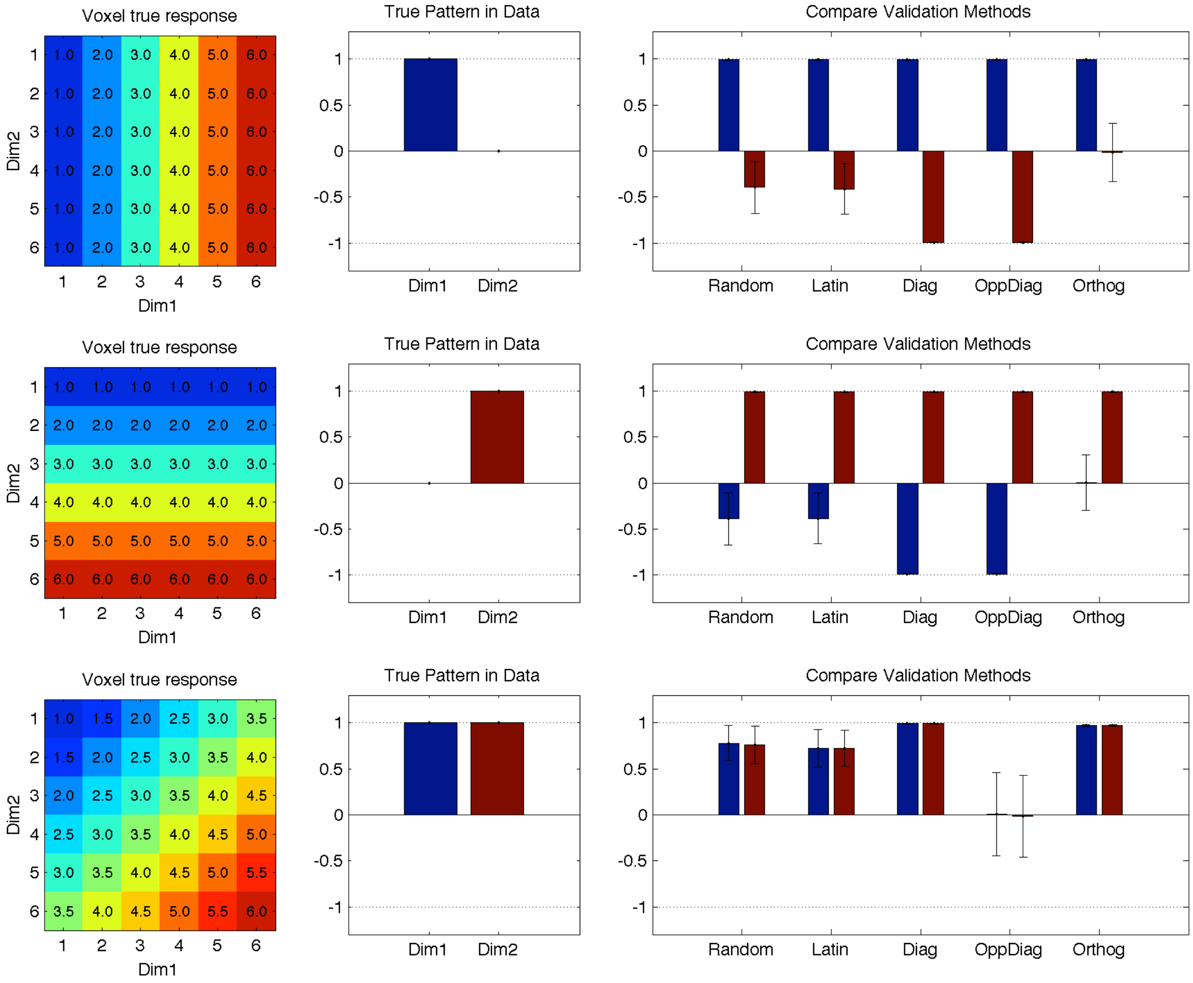

Figure S5. Summary of simulation results. Left Column: 3 example voxel data patterns. Middle Column: Voxel's true correlation with dimension1 or dimension2. Right Column: Mean correlations between the predicted and actual levels on the hold out data, averaged over 1000 iterations, for each of the hold-out validation procedures. Error bars reflect ± 1 standard deviation over iterations.