Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data

Fabian A. Soto[1], Lauren E. Vucovich[2], and F. G. Ashby[2]

[1]Department of Psychology, Florida International University, Modesto A. Maidique Campus, 11200 SW 8th St, Miami, FL 33199

[2]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106

Running title: Independent neural representation

Corresponding author:

Fabian A. Soto

Department of Psychology, Florida International University, Modesto A. Maidique Campus, 11200 SW 8th St, Miami, FL 33199

E-mail: fabian.soto@fiu.edu

Acknowledgements: This work was supported in part by grant no. W911NF-07-1-0072 from the U.S. Army Research Office through the Institute for Collaborative Biotechnologies, and by NIH grant 2R01MH063760.

Abstract

Many research questions in visual perception involve determining whether stimulus properties are represented and processed independently. In visual neuroscience, there is great interest in determining whether important object dimensions are represented independently in the brain. For example, theories of face recognition have proposed either completely or partially independent processing of identity and emotional expression. Unfortunately, most previous research has only vaguely defined what is meant by "independence," which hinders its precise quantification and testing. This article develops a new quantitative framework that links signal detection theory from psychophysics and encoding models from computational neuroscience, focusing on a special form of independence defined in the psychophysics literature: perceptual separability. The new theory allowed us, for the first time, to precisely define separability of neural representations and to theoretically link behavioral and brain measures of separability. The framework formally specifies the relation between these different levels of perceptual and brain representation, providing the tools for a truly integrative research approach. In particular, the theory identifies exactly what valid inferences can be made about independent encoding of stimulus dimensions from the results of multivariate analyses of neuroimaging data and psychophysical studies. In addition, two commonly used operational tests of independence are re-interpreted within this new theoretical framework, providing insights on their correct use and interpretation. Finally, we apply this new framework to the study of separability of brain representations of face identity and emotional expression (neutral/sad) in a human fMRI study with male and female participants.

Significance Statement

A common question in vision research is whether certain stimulus properties, like face identity and expression, are represented and processed independently. We develop a theoretical framework that allowed us, for the first time, to link behavioral and brain measures of independence. Unlike previous approaches, our framework formally specifies the relation between these different levels of perceptual and brain representation, providing the tools for a truly integrative research approach in the study of independence. This allows to identify what kind of inferences can be made about brain representations from multivariate analyses of neuroimaging data or psychophysical studies. We apply this framework to the study of independent processing of face identity and expression.

Introduction

A common goal in perceptual science is to determine whether some stimulus dimensions or components are processed and represented independently from other types of information. In visual neuroscience, much research has focused on determining whether there is independent processing of object and spatial visual information (Ungerleider and Haxby, 1994), of object shape and viewpoint (Rust and Stocker, 2010), of different face dimensions (Bruce and Young, 1986; Haxby et al., 2000), etc. A common approach is to use operational definitions of independence, which are linked to rather vague conceptual definitions. This approach has the disadvantage that different researchers use different operational definitions for independence, often leading to contradictory conclusions. For example, in the study of whether face identity and emotional expression are processed independently, evidence for both independence and interactivity has been found across a variety of operational tests. Evidence for independence was found by most lesion studies (Bate and Bennetts, 2015), by lack of correlation between fMRI patterns related to identity and expression (Hadj-Bouziane et al., 2008), by single neuron invariance (Hasselmo et al., 1989), by selective fMRI adaptation release in fusiform face area (FFA) and middle superior temporal sulcus (STS) (Winston et al., 2004), and by selective fMRI decoding of identity from anterior FFA and medial temporal gyrus (Nestor et al., 2011; Zhang et al., 2016), and of expression from STS (Zhang et al., 2016). Evidence for a lack of independence has been provided by overlapping fMRI activation during filtering tasks (Ganel et al., 2005), by non-selective fMRI adaptation release in posterior STS (Winston et al., 2004) and in FFA–when adaptation is calculated based on perception (Fox et al., 2009)–, and by non-selective fMRI decoding from right FFA (Nestor et al., 2011).

Because the different operational definitions are not linked to one another through a theoretical framework, the interpretation of such contradictory results is very difficult and necessarily post-hoc. Even more difficult is to link the neurobiological results to the psychophysics literature on independence of face dimensions, which itself is plagued by similar issues (for a review, see Soto et al., 2015).

General recognition theory (GRT, Ashby and Townsend, 1986; Ashby and Soto, 2015) is a multidimensional extension of signal detection theory that has solved such problems in psychophysics, by providing a unified theoretical framework in which notions of independence can be defined and linked to operational tests. Hundreds of studies have applied GRT to a wide variety of phenomena, including face perception (Thomas, 2001; Wenger and Ingvalson, 2002), recognition and source memory (Banks, 2000; Rotello et al., 2004), source monitoring (DeCarlo, 2003), object recognition (Cohen, 1997; Demeyer et al., 2007), perception/action interactions (Amazeen and DaSilva, 2005), speech perception (Silbert, 2012), haptic perception (Giordano et al., 2012), the perception of sexual interest (Farris et al., 2010), and many others.

Here we present an extension of GRT to the study of independence of brain representations, by relating it to encoding models and decoding methods from computational neuroscience (Ashby and Soto, 2016; Pouget et al., 2003). Past neuroimaging studies have been limited to choosing between decoding methods, which try to determine what stimulus information is processed in a brain region while ignoring the form of the underlying representation, and encoding models, which assume a specific representation and compare its predictions against data. (Naselaris et al., 2011). We propose the concept of encoding separability as a fundamental way in which brain representations of stimulus properties can be considered independent, and we identify the specific conditions in which a decoding analysis of neuroimaging data or a psychophysical study allow inferences to be made about encoding separability. In doing so, we show that decoding methods (and under some assumptions, psychophysics) can be useful to make valid inferences about encoding. We also re-interpret previously-proposed tests of independence within our new theoretical framework, and provide guides on their correct use. Finally, we apply this new framework to the study of separability of brain representations of face identity and expression.

Materials and Methods

We recommend that readers skip directly to the Results section and read all the theoretical results first, and come back to this section after reaching the section "An Application to the Study of Encoding Separability of Face Identity and Expression."

Participants

Twenty-one male and female right-handed students at the University of California Santa Barbara were recruited to participate in this study. Each participant received a monetary compensation at a rate of US$20/hour. This study was approved by the Human Subjects Committee at the University of California, Santa Barbara, and written informed consent was obtained from all participants.

Experimental Task

The stimuli and task were identical to those used in a previous behavioral study of separability of face identity and expression (Soto et al., 2015). Stimuli were four grayscale images of male faces, part of the California Facial Expression database (http://cseweb.ucsd.edu/~gary/CAFE/). Each face showed one of two identities with either a neutral or sad emotional expression. The faces were shown through an elliptical aperture in a homogeneous gray screen; this presentation revealed only inner facial features and hid non-facial information, such as hairstyle and color.

Participants performed an identification task both outside and inside the MRI scanner. Each stimulus was assigned to a specific response key and the participant's task was to identify the image presented in each trial. Each trial started with the presentation of a white crosshair in the middle of the screen for 200 ms, followed by stimulus presentation for a single frame (i.e., 16.667 ms at a 60 Hz refreshing rate). Stimulus presentation was short to make it identical to that used in our previous behavioral study. After stimulus presentation, participants pressed a response key; 500 ms later, feedback about the correctness of their response was displayed on the screen for 500 ms ("Correct" in green font color or "Incorrect" in red font color). If the participant took longer than 5 s to respond, the words "Too Slow" were presented on the screen and the trial stopped. Feedback was followed by a variable inter-trial interval, obtained by randomly sampling a value from a geometric distribution with parameter 0.5 and truncated at 5, and multiplying that value by 1,530 ms (the TR value, see below). To obtain estimates of stimulus-related activity with other events in the trial (crosshair and response) unmixed, we used a partial trials design in which 25% of the trials included the presentation of the white crosshair not followed by a stimulus. Participants were instructed to randomly choose a response on these partial trials.

Stimulus presentation, feedback and response recording were controlled using MATLAB augmented with the Psychophysics Toolbox (psychtoolbox.org), running on Mackintosh computers. Participants practiced the identification task on personal Mackintosh computers outside the MRI scanner for about 20 mins. During this training, participants responded on a keyboard. During scanning, participants responded using the Lumina Response Pad System (model LU400-Pair), with the same finger-stimulus mapping as during pre-training.

Functional Imaging

Images were obtained using a 3T Siemens TIM TRIO MRI scanner with a 12-channel head coil at the University of California, Santa Barbara Brain Imaging Center. Cushions were placed around the head to minimize head motion. A T1-weighted high-resolution anatomical scan was acquired using an MPRAGE sequence (TR: 2,300 ms; TE: 2.98 ms; FA: 9°; 160 sagittal slices; 1 × 1 × 1 mm voxel size; FOV: 256 mm). Additional scans included a localizer and a GRE field map, neither of which were used in the analyses presented here.

Functional scans used a T2*-weighted single shot gradient echo, echo-planar sequence sensitive to BOLD contrast (TR: 1,530 ms; TE: 28 ms; FA: 61°; FOV: 192 mm) with generalized auto-calibrating partially parallel acquisitions (GRAPPA). Each volume consisted of 28 slices (interleaved acquisition, 2.5 mm thick with a 0.5 mm gap; 2.5 × 2.5 mm in-plane resolution) acquired at a near-axial orientation, manually adjusted

to cover the ventral visual stream and lateral prefrontal cortex. There were a total of four functional runs per participant (with the exception of five participants who completed three functional runs).

The first run was a standard functional localizer for face regions (Fox et al., 2009). Neutral faces, emotional faces and non-face objects were each presented in different stimulus blocks, separated by fixation blocks. Sixteen images of the same type were presented within a stimulus block, each with a duration of 500 ms and a 250 ms inter-stimulus-interval. Fixation blocks consisted of the presentation of a black screen with a white fixation cross in the middle. The sequence started with a fixation block, followed by 6 blocks of each image category (18 total), each followed by a fixation block, for a total of 37 blocks. Blocks lasted for 12 seconds, and the whole scan lasted about 7.5 mins. The order of image types (e.g., neutral-emotional-object) was counterbalanced across blocks. To ensure attention to the stimuli, participants were asked to push a button whenever an image was repeated in the sequence. Four of the 15 stimuli in a block were repetitions, randomly positioned in the stimulus sequence.

In all other functional runs, which lasted about 10 mins each, participants performed the identification task described earlier, without feedback. Each of the four images was repeated 25 times, for a total of 100 trials per run. Stimuli were viewed through a mirror mounted on the head coil and a back projection screen.

Statistical Analyses

Anatomical scans

Processing of structural scans was done using FSL (www.fmrib.ox.ac.uk/fsl), and included brain extraction using BET and nonlinear registration to MNI 2mm standard space using FNIRT nonlinear registration. The inverse transformation was obtained to transform volumes from standard space back to subject space. For visualization purposes, some statistical maps were converted from MNI standard space to the PALS-B12 surface-based atlas using CARET v.5.65 (www.nitrc.org/projects/caret/) and selecting the options *average of the mapping to all multi-fiducial cases* and *enclosing voxel algorithm.*

Preprocessing of functional scans

Preprocessing of the functional scans was conducted using FEAT (fMRI Expert Analysis Tool) version 6.00, part of FSL (www.fmrib.ox.ac.uk/fsl). Volumes from all three runs of the main identification task were concatenated into a single series using *fslmerge*. Preprocessing included motion correction using MCFLIRT, slice timing correction (via Fourier time-series phase-shifting), BET brain extraction, grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor, and a high-pass temporal filtering

(Gaussian-weighted least-squares straight line fitting, with sigma=50.0s). The data from the functional localizer were spatially smoothed with a Gaussian kernel of FWHM 4.0mm. The data used in the main separability analysis were not spatially smoothed during preprocessing. Each functional scan was registered to the corresponding structural scan using boundary-based registration (BBR) in FLIRT with default parameters.

Neural activity estimates

After preprocessing of the functional scans, estimates of single-trial stimulus-related activity were obtained for the faces in the main identification task. We used the iterative FBR (finite BOLD response) method described by Turner et al. (Turner et al., 2012) to deconvolve the BOLD activity related to each stimulus presentation. This method avoids assumptions about the hemodynamic response function that are inherent to parametric estimation methods and it is more successful than the latter in unmixing the responses to temporally adjacent events in event-related designs (Turner et al., 2012). Instead of assuming a particular shape of the hemodynamic response function, the full shape of the BOLD response to a stimulus is estimated through a set of 12 FBR regressors that are ordered in sequence. In the regression matrix, each event is represented by a set of 12 ones, starting at the beginning of the event. The method is called "iterative" because it iterates through each trial to estimate the BOLD activity related to the stimulus presentation in that trial only. To do this, a group of 12 regressors is created for the target stimulus, and separate groups of 12 regressors are created for the four stimulus classes in the experiment (the target trial was excluded from the regressor of its class), and for the conjunction of crosshair presentation and response. This results in the estimation of 12 regression coefficients for the target stimulus, which are kept while all other regressors are discarded (they are included only to unmix their influence from the target estimates of the BOLD response). As indicated above, the process is iterated for each trial, resulting in a set of spatiotemporal maps (one for each stimulus presentation), representing estimates of the BOLD activity in each voxel and each of 12 TRs starting at the time of stimulus presentation. The algorithm was implemented in MATLAB (The MathWorks, Natick, MA, USA).

Decoding separability test

Here we describe the procedures used to implement a decoding separability test. Theoretical results linking this test to the notions of decoding separability, encoding separability and perceptual separability, as well as a justification for the application of this test to neuroimaging data, can be found in the Results section.

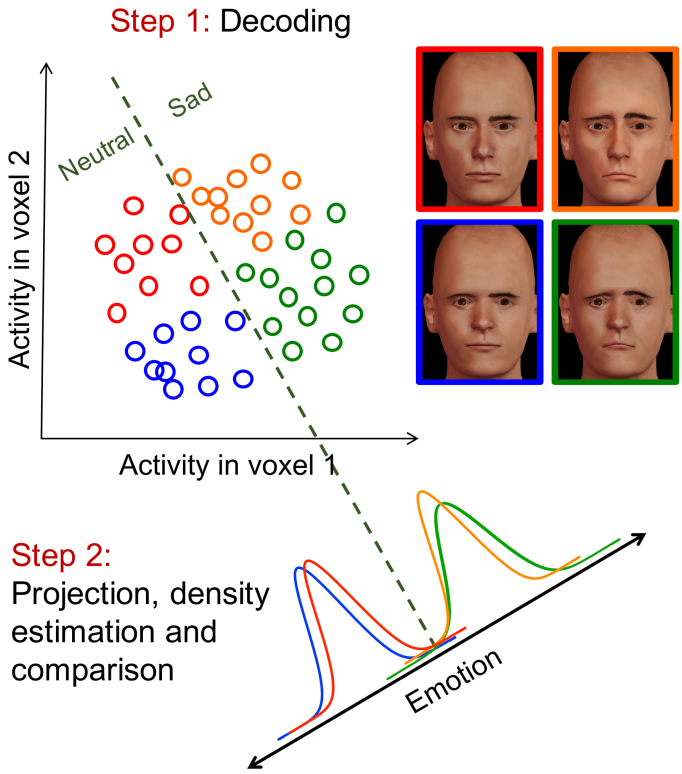

Figure 1: A schematic representation of a test of decoding separability for neuroimaging data, implemented as an extension to traditional linear decoding procedures. The simplified example considers the representation of four stimuli in two voxels. Each point represents activity on a different trial, and each color represents a different stimulus that has been repeatedly presented during the experiment. The dotted line represents a classification bound that separates trials according to emotional expression. The line orthogonal to this bound represents the direction in voxel space that best discriminates one expression from the other. Decoding separability holds if the distributions along this dimension for a given value of the target dimension (emotional expression) are equivalent across changes in the irrelevant dimension (identity). Adapted from: http://figshare.com/articles/Test-of-separabilityof-neural-representations/1385406

Figure 1 is a schematic representation of the decoding separability test. In this simplified example, we consider only two voxels. The estimates of activity are thus represented in a two-dimensional voxel space. Each point represents activity on a different trial, with each color representing a different stimulus that has been repeatedly presented during the experiment. Decoding facial expression from these two voxels using a linear classifier involves finding a hyperplane in the activity space that best separates trials in which a neutral face was shown from trials in which a sad face was shown (the dotted line in Figure 1). The line orthogonal to the classification bound (sometimes called the classifier's "decision variable") represents the direction in voxel space that best discriminates one expression from the other. Thus, it is reasonable to assume that this is the direction in this specific voxel space along which expression is encoded. Using that specific direction in voxel space is not a requirement of the decoding separability test to be valid (the only

requirement is that a linear decoding scheme is used; see Results section), but it allows us to link the present work to more traditional MVPA techniques. If we take all the observed data points and project them onto this "expression" dimension, we can use these projected points to estimate a distribution of decoded values. *Decoding separability* holds if this distribution of decoded values is invariant across changes in the stimulus on a second, irrelevant dimension. To test for decoding separability, the two distributions of points for a given expression (e.g., the orange and green distributions for "sad"), each corresponding to a different identity, can be compared to one another.

In the Results section, we link this decoding separability test to our main theory, and show that it is a valid test of violations of separability in neural representations, even when it is applied to indirect and noisy measures of brain activity, like those obtained from fMRI.

In two separate analyses, we tested the separability of emotion from identity and the separability of identity from emotion; because both analyses are identical, we will describe the analysis in terms of decoding separability of a "target" dimension from an "irrelevant" dimension. Each of the regression coefficients obtained in the previous step were standardized by column. We used a searchlight procedure (Kriegeskorte et al., 2006) with a spherical mask that had a radius of three voxels. In each step of the analysis, the searchlight was centered on a different brain voxel and the selected data were used to train a linear support vector machine (SVM, using the *LinearNuSVMC* classifier included in pyMVPA) to decode the target dimension using all the available data. Then the data were projected to the normal line to the classification hyperplane to obtain a number of decoded values on the target dimension. Using Python augmented with the SciPy ecosystem, the group of decoded values for each stimulus was used to obtain kernel density estimates (KDEs) of their underlying probability distribution. A gaussian kernel and automatic bandwidth determination were used as implemented in the SciPy function *gaussian_kde*. Let $\hat{p}_{ij}\left(\hat{A}\right)$ represent the KDE for a stimulus with value i on the target dimension and value j on the irrelevant dimension, evaluated at point $\hat{A}$. The index i can take one of two values representing, for example, "sad" and "neutral" when the target dimension is emotional expression, as in the example given in Figure 1. Similarly, the index j can take one of two values representing "identity 1" and "identity 2", as in this example identity is the irrelevant dimension. Then an index of deviations from decoding separability (DDS) was computed from all four KDEs obtained from the target dimension (in this example, emotional expression), according to the following equation:

$$DDS = \sum_{i=1}^{2} \sum_{k=1}^{1{,}000} \left| \hat{p}_{i1}\left(\hat{A}_k\right) - \hat{p}_{i2}\left(\hat{A}_k\right) \right| \tag{1}$$

Each KDE was evaluated at 1,000 evenly-spaced points $\hat{A}_k$, indexed by $k = 1, 2...1,000$, starting at the

minimum data point minus half the data range, and finishing at the maximum data point plus half the data range. Note that the value $\hat{p}_{i1}\left(\hat{A}_k\right) - \hat{p}_{i2}\left(\hat{A}_k\right)$ represents the difference between two distributions of decoded values, both related to stimuli with the same value in the target dimension (e.g., "sad", represented by the index i) but different values in the irrelevant dimension (e.g., "identity 1" and "identity 2", represented by the indexes 1 and 2 in the equation). The index uses the absolute value of the difference between a pair of distributions, which by definition is the $L1$ distance between the two (discretized) distributions (see Equation 25 below). If separability holds, then the distance between distributions should be zero. However, this is only true if we had access to the true distributions. Any error in the KDEs should produce differences between distributions that are added to the DDS. This makes it difficult to statistically test for deviations of separability, as the data from multiple participants cannot be combined (differences in the estimation error of the KDEs produces differences in scale of the statistic) and the expected value of the statistic under the null hypothesis is unknown. Under the assumption of decoding separability, two distributions that share the same level of the target dimension but different levels of the irrelevant dimension are identical. That is, data points from those distributions are exchangeable. Taking this into account, we standardized the statistic in the following way: (1) we shuffled the level of the irrelevant dimension for each data point 200 times (separately for each level of the target dimension); (2) each time we computed the DDS, yielding an empirical distribution function (EDF) of the statistic under the assumption of decoding separability; (3) the final standardized value was the percentile of the observed DDS in the EDF minus 50, representing percentile deviation from the median of the EDF.

Repeating this process for all searchlights resulted in a DDS map for each participant, which were converted to the participant's anatomical space using FSL's FLIRT linear registration and then to MNI 2mm standard space using FSL's FNIRT nonlinear registration. The resulting DDS maps in standard space were submitted to a nonparametric permutation test using FSL's *randomise* program (Winkler et al., 2014), with the option *clusterm* for correction for multiple comparisons (which uses the distribution of the maximum cluster mass in the permutation test), a cluster threshold of 2.53 (corresponding to p=0.01, uncorrected), variance smoothing with a sigma of 5 mm, and 5,000 permutations.

For visualization purposes, the volumes with significant statistics obtained from the permutation test were converted to the PALS-B12 surface-based atlas using CARET v.5.65 (www.nitrc.org/projects/caret/), and displayed together with the borders of face-selective areas from the localizer scan.

Face-selective regions

Face-selective regions were defined using the data from the functional localizer. Low-level analyses were performed separately on the data from each participant. Three explanatory variables (EVs) were defined:

Neutral Faces, Emotional Faces and Objects, each corresponding to a boxcar function covering the corresponding blocks in the functional scan (see Functional Localizer description above). These boxcar functions were convolved with the default Gamma hemodynamic response function in FSL, which has a mean lag of 6s and a standard deviation of 3s. A temporal derivative and temporal filtering were added to the design matrix. Two contrasts were formed: Faces (Neutral Faces + Emotional Faces) > Objects, to define regions selective to face information in general, and Emotional Faces > Neutral Faces, to define regions selective to face emotional expression more specifically. Each of these contrasts resulted on a separate map of z statistics for each participant. The individual z statistical maps were used as input to a high-level analysis, using a mixed-effects model (the option FLAME 1+2 in FSL), to generate a group map for each contrast. Clusters were first identified by thresholding the maps at z=2.3; the experiment-wise false positive rate ($\alpha = 0.05$) was controlled by using a threshold on cluster size derived from Gaussian random field theory.

The volumes with significant clusters obtained from the two contrasts were converted to the PALS-B12 surface-based atlas and their borders were manually drawn using CARET v.5.65 (http://www.nitrc.org/projects/caret/). These region borders were used as rough landmarks for the interpretation of the main results of the decoding separability analysis.

The Faces > Objects contrast was additionally used to define face-selective functional regions of interest (ROIs) using the Group-Constrained Subject-Specific (GSS) described by (Julian et al., 2012). First, individual maps were thresholded at $p < 0.05$, uncorrected, and the resulting thresholded images were binarized. It was necessary to use a much more liberal threshold than that used by Julian et al. ($p < 0.0001$) to obtain ROI masks in most participants (even at this low threshold, we did not obtain an ROI for the OFA in one participant), because our study was designed to carry out analyses at the group level (see below) and therefore the contrast had less power than that of Julian et al. at the level of individual participants. Second, we took the group-level "parcels" provided by Julian et al. in MNI 2mm standard space (available at http://web.mit.edu/bcs/nklab/GSS.shtml), and transformed them to the participant's functional space using FNIRT. Third, we intersected the individual binary maps and the group-level parcels to define ROIs corresponding to the fusiform face area (FFA), occipital face area (OFA) and superior temporal sulcus face area (STS) in each individual participant.

Additional anatomical ROIs were obtained, to serve as controls and explore the behavior of our decoding separability test in different conditions. We obtained an ROI corresponding to primary visual cortex (PVC) from the Juelich Histological Atlas, and an ROI corresponding to the lateral ventricles from the Harvard-Oxfor Subcortical Structural Atlas; both atlas are included with FSL. The obtained ROIs were thresholded at a value of 20 using *fslmaths* and binarized. These final ROI masks, which were in MNI 2mm standard space, were transformed to each participant's functional space using FNIRT.

All ROIs were obtained for both the left and right hemispheres.

Results

Extending General Recognition Theory to the Study of Brain Representations

GRT is a multivariate extension of signal detection theory to cases in which stimuli vary on more than one dimension (Ashby and Townsend, 1986; Ashby and Soto, 2015). As in signal detection theory, the theory assumes that different presentations of the same stimulus produce slightly different perceptual representations. For example, as shown in Figure 2, repeated presentations of a face identity produce a variety of values on the "identity" dimension (blue and red dots), which follow a probability distribution (red and blue curves). According to GRT, there are many ways in which processing of a dimension of interest, or target dimension, can be influenced by variations in a second, irrelevant dimension. GRT formally defines such dimensional interactions and links them to operational tests of independence. This allows researchers to determine whether a test can dissociate between different forms of independence, and to create new tests specifically designed to target a specific form of independence.

Here we will consider the special case in which stimuli vary along two stimulus dimensions (or more generally, components or properties), represented by A and B. However, the theory can easily be extended to a larger number of dimensions. Specific values of dimension A used in an experiment are indexed by $i = 1, 2, ...L_A$, and the specific values of dimension B are indexed by $j = 1, 2, ...L_B$. A stimulus in the experiment is represented by a combination of these dimension levels, A_iB_j. This stimulus produces a random perceptual effect in a two-dimensional perceptual space $[x, y]$, where x represents the perceptual effect of property A and y the perceptual effect of property B. The random vector $[x, y]$ can be described through a two-dimensional joint probability density $p\,(x, y|A_iB_j)$, with $p\,(x|A_iB_j)$ and $p\,(y|A_iB_j)$ representing the marginal densities of the perceptual effects associated with components A and B, respectively (the distributions shown in Figure 2 are examples of such marginal densities).

Perceptual separability

A particularly important form of independence defined in GRT is *perceptual separability*, which holds when the perception of the target dimension is not affected by variations in the irrelevant dimension. In Figure 2, an identity is presented with a neutral expression (in blue) or with a sad expression (in red). When perceptual separability holds, the blue and red perceptual distributions overlap, and the face is just as easy to identify in both cases. When perceptual separability fails, the blue and red perceptual distributions do

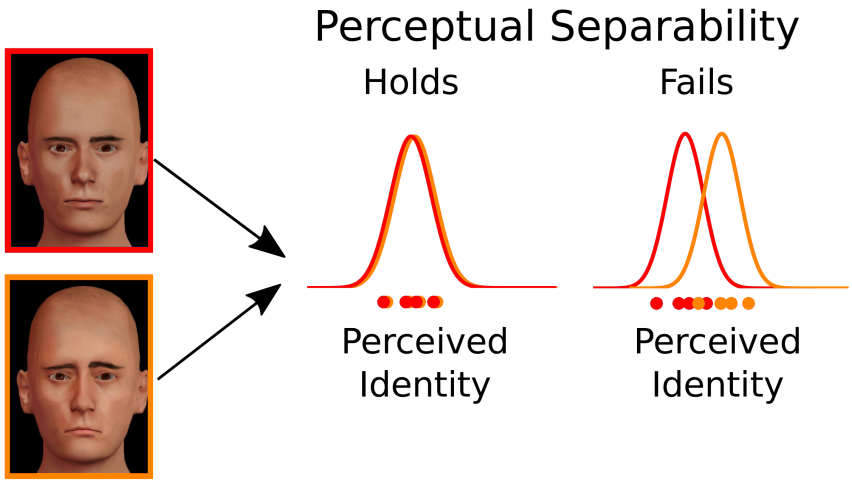

Figure 2: Stimulus representation and definition of perceptual separability in GRT. The representation of a given identity changes randomly from trial to trial (dots at the bottom) according to some perceptual distribution (bell-shaped distributions at the top). Perceptual separability of identity from emotional expression (neutral vs. sad) holds if the perceptual distribution for identity does not change with emotional expression (left), and it fails if the perceptual distribution for identity does change with emotional expression (right).

not overlap, and the face is easier to identify when the expression is sad (there is more evidence for the identity in this case).

Here we focus on perceptual separability because it is considered a particularly important form of independence, for two reasons. First, because many questions in perceptual neuroscience can be understood as questions about separability of object dimensions. For example, the question of whether object representations are invariant across changes in identity-preserving variables like rotation and translation is equivalent to the question of whether object representations are perceptually separable from such variables (Stankiewicz, 2002). In face perception, configural or holistic face perception has been defined as non-separable processing of different face features (Mestry et al., 2012; Richler et al., 2008), and the question of whether or not different face dimensions are processed independently is usually investigated using tests of perceptual separability (Fitousi and Wenger, 2013; Ganel and Goshen-Gottstein, 2004; Schweinberger and Soukup, 1998; Soto et al., 2015). The second reason for the importance of perceptual separability is that higher-level cognitive mechanisms seem to be applied differently when stimuli differ along separable dimensions rather than along non-separable dimensions. For example, selective attention is deployed more easily to separable dimensions than to non-separable dimensions (Garner, 1974; Goldstone, 1994), sources of predictive and causal knowledge may be combined differently if they differ along separable versus non-separable dimensions (Soto et al., 2014, 2015), and the performance cost of storing objects in visual working memory is different depending on whether such objects differ from one another in separable versus non-separable dimensions (Bae and

Flombaum, 2013).

Formally, perceptual separability of dimension A from dimension B occurs when the perceptual effect of stimuli on dimension A does not change with the value of the stimulus on dimension B (Ashby and Townsend, 1986)–that is, if and only if, for all values of x and i:

$$p\left(x|A_iB_1\right) = p\left(x|A_iB_2\right) ... = p\left(x|A_iB_{L_B}\right). \tag{2}$$

Perceptual separability of dimension B from dimension A is defined analogously.

Neural encoding and encoding separability

Extending GRT to the study of neural representation requires linking it to our current understanding on how dimensions are represented by neuronal populations. In the computational neuroscience literature, an encoding model is a formal representation of the relation between sensory stimuli and the response of a single neuron or a group of neurons (Pouget et al., 2000, 2003; Ma, 2010). In the case of stimulus dimensions, an encoding model represents how changes in a dimension of interest are related to changes in neural responses. Encoding models have been applied to describe neural responses at a variety of scales, from single neurons to the average activity of thousands of neurons (Brouwer and Heeger, 2009; Pouget et al., 2000, 2003; Naselaris et al., 2011). To discuss these models in their more general form, it is convenient to introduce the abstract concept of a *channel*, which can be used as a placeholder for a single neuron, a population of neurons with similar properties, or as an abstract construct to model the behavior of a human observer. A channel is essentially a detector, sensitive to a particular stimulation pattern. It responds maximally to that target pattern and progressively less to other patterns as they become different from the target. In other words, the most important property of a channel is that it has tuning. The tuning of a channel can be modeled in many ways, but perhaps the simplest is to choose a physical dimension of interest and model the channel's response as a function of the value of a stimulus on that dimension. For example, if we are interested in dimension A (equivalent definitions can be given for B), then the response r_c of the channel c to a stimulus A_iB_j is determined by a tuning function:

$$r_c\left(A_iB_j\right) = f_c\left(A_iB_j\right). \tag{3}$$

Common choices for $f_c\left(A_iB_j\right)$ in the literature are bell-shaped and sigmoidal functions (Pouget et al., 2003). The channel response on a given trial may also be influenced by stochastic internal noise, which can be assumed to be additive (independent of the channel's response) or multiplicative (scaling with the channel's response). Common choices for the distribution of this noise in the literature are Gaussian and

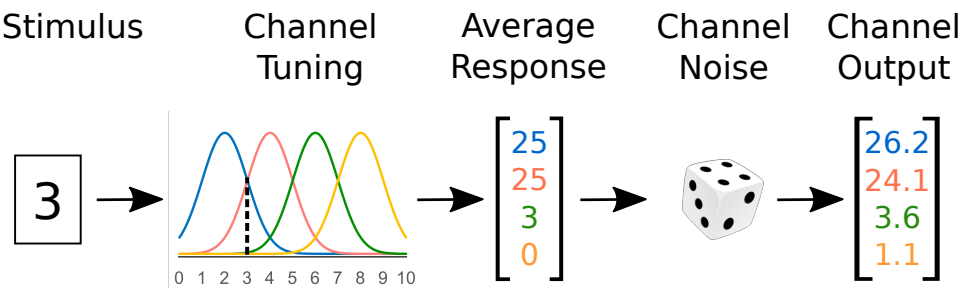

Figure 3: Schematic representation of multiple channels encoding a stimulus with a value of "3" in a target dimension. If a stimulus with value "3" is presented, each channel gives an average response equivalent to the height of the tuning function at that stimulus value (i.e., the height at the dotted line). The vector of average responses is perturbed by random noise, producing the final channel output.

Poisson (Pouget et al., 2003; Ma, 2010). Because the noise is a random variable, the response of the channel r_c itself becomes a random variable that follows a probability distribution:

$$p\left(r_c|A_iB_j,\theta\right) \sim \eta\left(f_c\left(A_iB_j\right),\theta\right), \tag{4}$$

where $\sim$ means "distributed as", and $\eta\left(\right)$ is just a placeholder that stands for any probability distribution (e.g., Gaussian) that depends on the channel's tuning function and on a set of parameters θ describing noise.

Researchers agree that encoding of a stimulus dimension requires a model with multiple channels, or multichannel model. For example, the "standard model" of dimension encoding in the computational neuroscience literature is such a multichannel model (implementing a "population code" (Pouget et al., 2000, 2003)), and most applications in the neuroscience and psychophysics literature use at least two channels to describe encoding of stimulus dimensions (Gold and Shadlen, 2001). Figure 3 shows encoding of a stimulus dimension with four channels, each with its own tuning model represented by a curve of different color. The tuning model is a formalization of how the channel responds to different stimulus values: each channel responds maximally to its preferred dimensional value and less to other values. The figure shows the response of a multi-channel model to a stimulus with a value of 3 on the target dimension. A channel's noise model describes the stochasticity in the channel's responses through a probability distribution. In Figure 3, the average response of each channel is perturbed by random additive noise, represented by the dice. The final channel output is equal to the average response (from the tuning model) plus noise (randomly drawn from the noise model).

A multichannel model encodes information about dimension A through the combined response of N channels (Pouget et al., 2000, 2003; Gold and Shadlen, 2001), indexed by $c = 1, 2, ...N$. On each trial, the model produces a (column) random vector of channel responses $\mathbf{r} = [r_1, r_2, ... r_N]^\intercal$, where $^\intercal$ denotes matrix

transpose. Note that $\mathbf{r}$ depends on the stimulus A_iB_j according to a set of tuning functions:

$$\mathbf{f}\left(A_iB_j\right) = \left[f_1\left(A_iB_j\right), f_2\left(A_iB_j\right), ...f_N\left(A_iB_j\right)\right]^\intercal, \tag{5}$$

The probability distribution of $\mathbf{r}$ also depends on noise parameters θ:

$$p\left(\mathbf{r}|A_iB_j,\theta\right) \sim \eta\left(\mathbf{f},\theta\right). \tag{6}$$

As indicated earlier, additive Gaussian noise is a common choice for the channel noise model. In that case, the multichannel encoding model is described by a multivariate Gaussian distribution:

$$p\left(\mathbf{r}|A_iB_j\right) \sim \mathcal{N}\left(\mathbf{f}\left(A_iB_j\right), \mathbf{\Sigma}\left(A_iB_j\right)\right), \tag{7}$$

where $\mathbf{\Sigma}\left(A_iB_j\right)$ is an $N \times N$ covariance matrix describing channel noise. In most applications, noise is also assumed to be independently distributed across channels, and all non-diagonal cells in $\mathbf{\Sigma}\left(A_iB_j\right)$ are zero.

This discussion suggests that a new form of separability can be defined for the neural representation of a dimension: *encoding separability*. When a target dimension is encoded in the exact same way across variations of an irrelevant dimension, we say that the former shows encoding separability from the latter. For encoding separability to hold, both the tuning and noise models of all channels must be equivalent across changes in the irrelevant dimension, which is equivalent to having a single encoding model representing the target dimension, independently of the value of the irrelevant dimension.

Formally, encoding separability of dimension A from dimension B holds when encoding of the value of A does not change with the stimulus' value on B. That is, if and only if, for all values of $\mathbf{r}$ and i:

$$p\left(\mathbf{r}|A_iB_1,\theta\right) = p\left(\mathbf{r}|A_iB_2,\theta\right) ... = p\left(\mathbf{r}|A_iB_{L_B},\theta\right). \tag{8}$$

Encoding separability of dimension B from dimension A is defined analogously.

Violations of encoding separability can happen for two reasons. The first possibility is that one or more of the tuning functions in $\mathbf{f}$ change with the value of B. *Tuning separability* of dimension A from dimension B holds when all tuning functions that encode dimension A depend only on the value of A–that is, if and only if, for all channels c and stimuli A_iB_j:

$$f_c\left(A_iB_j\right) = f_c\left(A_i\right). \tag{9}$$

Tuning separability of dimension B from dimension A is defined analogously. Because $p(\mathbf{r}|A_iB_j,\theta)$ depends on $\mathbf{f}$ (see Equation 6), violations of tuning separability produce violations of encoding separability.

The second reason for a violation of encoding separability is that the noise for one or more channels is distributed differently for different levels of B.

Because the Gaussian encoding model described in Equation 7 is completely characterized by the mean vector $\mathbf{f}(A_iB_j)$ and the covariance matrix $\mathbf{\Sigma}(A_iB_j)$, encoding separability of A from B holds if the following two conditions are true for all stimuli A_iB_j:

$$\begin{aligned}\mathbf{f}(A_iB_j) &= \mathbf{f}(A_i)\\ \mathbf{\Sigma}(A_iB_j) &= \mathbf{\Sigma}(A_i)\end{aligned} \tag{10}$$

Decoding separability

The term neural decoding refers both to a series of methods used by researchers to extract information about a stimulus from neural data (Naselaris et al., 2011; Quiroga and Panzeri, 2009) and to the mechanisms used by readout neurons to extract similar information, which is later used for decision making and other cognitive processes (Pouget et al., 2003; Seung and Sompolinsky, 1993). If dimension A is encoded by N channels, according to the scheme summarized in Equation 6 and depicted in Figure 3, then the decoded estimate of a dimensional value $\hat{A}$, will be some function of the channel responses:

$$\hat{A} = g(\mathbf{r}), \tag{11}$$

where $g()$ is a function from $\mathbb{R}^N$ to $\mathbb{R}$ (i.e., from the multidimensional space of the channel responses to the unidimensional space of the decoded dimension). Because $\mathbf{r}$ is a random vector (see Equation 6), the decoded value $\hat{A}$ is a random value that follows a probability distribution $p\left(\hat{A}|A_iB_j,\theta\right)$. In many cases, knowledge about the encoding distribution from Equation 6 and the decoder from Equation 11 allows one to derive an expression for $p\left(\hat{A}|A_iB_j,\theta\right)$.

There are many possible decoding schemes, but the most popular among researchers (Seung and Sompolinsky, 1993; Pereira et al., 2009), due to their simplicity and neurobiological plausibility, are simple linear decoders

$$\hat{A} = \beta + \mathbf{b}^{\intercal}\mathbf{r}, \tag{12}$$

where β is a scalar and $\mathbf{b}$ is a (column) vector of weights.

With a Gaussian encoding model like the one described by Equation 7, the distribution of linearly-decoded

estimates of values on dimension A is:

$$p\left(\hat{A}|A_iB_j\right) \sim \mathcal{N}\left(\beta + \mathbf{b}^\intercal \mathbf{f}\left(A_iB_j\right), \mathbf{b}^\intercal \mathbf{\Sigma}\left(A_iB_i\right)\mathbf{b}\right). \tag{13}$$

When channel noise is independent, the variance of the decoded variable $\hat{A}$ is simply $\sum_{k=1}^{N} b_k^2\sigma_k^2$, where σ_k^2 represents the N diagonal elements of $\mathbf{\Sigma}\left(A_iB_j\right)$.

We define *decoding separability* as the situation in which the distribution of decoded values on the target dimension is invariant across changes in the stimulus on a second, irrelevant dimension. That is, decoding separability of dimension A from dimension B holds when the distribution of decoded values of A does not change with the value of B in the stimulus–that is, if and only if, for all values of $\hat{A}$ and i:

$$p\left(\hat{A}|A_iB_1,\theta\right) = p\left(\hat{A}|A_iB_2,\theta\right) ... = p\left(\hat{A}|A_iB_{L_B},\theta\right). \tag{14}$$

Decoding separability of dimension B from dimension A is defined analogously.

Relation between encoding separability and decoding separability

Decoding separability is easy to check by directly decoding dimensional values from a neuronal population. Moreover, if the same decoding scheme is used for all values of the irrelevant dimension, then the relations between encoding separability and decoding separability shown in Figure 4 hold, as we show in this section.

If encoding separability holds, then decoding separability must also hold. This proposition is represented by the green arrow in Figure 4. When encoding separability holds (see Equation 8), $p\left(\mathbf{r}|A_iB_j,\theta\right) = p\left(\mathbf{r}|A_i,\theta\right)$ for all values of $\mathbf{r}$ and j. Because we have assumed that decoding depends only on the value of $\mathbf{r}$ (Equation 11), the function $g\left(\right)$ is also independent of the value of B_j. Thus, regardless of the shape of $p\left(\mathbf{r}|A_i,\theta\right)$ and $g\left(\right)$, the distribution of the decoded variable $\hat{A}$ is independent of the value of B_j, and decoding separability (Equation 14) holds. In other words, for all values of B_j the same decoding transformation $g\left(\right)$ is applied to the same encoding distribution $p\left(\mathbf{r}|A_i,\theta\right)$, resulting in the same decoding distribution $p\left(\hat{A}|A_i,\theta\right)$.

If encoding separability fails, then decoding separability may fail or hold. This is represented by the red arrows in Figure 4. Our strategy to prove this proposition will be to disprove two universal statements through counterexamples.

We start by offering a counterexample disproving the following universal statement: *if encoding separability fails, then decoding separability must fail*. Suppose that a dimension A is encoded through the model with Gaussian channel noise described by Equation 7, and that we use a linear decoder to estimate $\hat{A}$, as

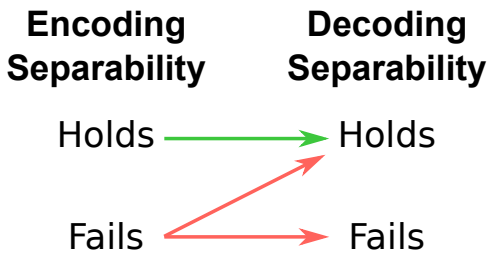

Figure 4: Summary of the relation between encoding separability and decoding separability, according to our extension to GRT. Arrows should be interpreted as conditional statements of the form "if X, then Y". These relations mean that a failure of encoding separability is a valid inference from the observation of a failure of decoding separability. However, the presence of encoding separability cannot be validly inferred from an observation of decoding separability.

described by Equation 12. Also suppose that there are violations of tuning separability (Equation 9) of A from B in the encoding model. Without loss of generality, suppose that those violations are differences in the tuning functions of A_1B_1 and A_1B_2:

$$\mathbf{f}\left(A_1B_1\right) \neq \mathbf{f}\left(A_1B_2\right),$$

or equivalently

$$\mathbf{f}\left(A_1B_1\right) = \mathbf{f}\left(A_1B_2\right) + \delta,$$

where δ represents a $N \times 1$ vector of deviations from tuning separability, and $\delta \neq \mathbf{0}$.

Under the assumptions listed above, the tuning functions only affect the mean of the decoded variable (see Equation 13), so we can ignore its variance. Now suppose that in this model decoding separability holds. In that case, we have that:

$$\begin{aligned} \beta + \mathbf{b}^\intercal \mathbf{f}\left(A_1B_1\right) &= \beta + \mathbf{b}^\intercal \mathbf{f}\left(A_1B_2\right) \\ \mathbf{b}^\intercal\left(\mathbf{f}\left(A_1B_2\right) + \delta\right) &= \mathbf{b}^\intercal \mathbf{f}\left(A_1B_2\right) \\ \mathbf{b}^\intercal \delta &= 0 \end{aligned} \tag{15}$$

For any given $\delta \neq \mathbf{0}$, there are an infinite number of $\mathbf{b} \neq \mathbf{0}$ that satisfy this equation, yielding a model in which encoding separability fails and decoding separability holds. The universal statement *if encoding separability fails, then decoding separability must fail* is false.

We now offer a counterexample to disprove the alternate universal statement: *if encoding separability*

fails, then decoding separability must hold. Following the same line of reasoning as before, we get that if decoding separability fails then:

$$\begin{aligned} \beta + \mathbf{b}^\intercal \mathbf{f}\left(A_1 B_1\right) &= \beta + \mathbf{b}^\intercal \mathbf{f}\left(A_1 B_2\right) + d \\ \mathbf{b}^\intercal \left[\mathbf{f}\left(A_1 B_2\right) + \delta\right] &= \mathbf{b}^\intercal \mathbf{f}\left(A_1 B_2\right) + d \\ \mathbf{b}^\intercal \delta &= d, \end{aligned} \tag{16}$$

where d represents a scalar deviation from decoding separability in the mean of the decoding distributions. As before, for any given $\delta \neq \mathbf{0}$ and $d \neq 0$, there are an infinite number of $\mathbf{b} \neq \mathbf{0}$ that satisfy this equation, yielding a model in which encoding separability fails and decoding separability fails. The universal statement *if encoding separability fails, then decoding separability must hold* is false.

In summary, if encoding separability fails, then decoding separability may hold or fail. While no universal statements can be made about decoding separability when encoding separability fails, a more specific relation may hold for particular combinations of encoding models and decoding schemes. For example, it might be that for some specific combination of an encoding model and decoding scheme, a failure of encoding separability necessarily leads to a failure of decoding separability, eliminating the diagonal arrow in Figure 4. If that was the case, it would be possible to infer encoding separability from a finding of decoding separability. We will not explore these possibilities here and instead will leave them for future work. However, note that the counterexamples offered here involve normally-distributed channel noise and a linear decoder, both of which are common choices in the literature on encoding and decoding. That is, under common assumptions and methods it is not possible to infer encoding separability from a finding of decoding separability.

One general result that must hold true is the following: if encoding separability fails and $\hat{A} = g\left(\mathbf{r}\right)$ is an injective (one-to-one) mapping, then decoding separability must fail. This is the case because when encoding separability fails (without loss of generality) $p\left(\mathbf{r}|A_1 B_1, \theta\right) \neq p\left(\mathbf{r}|A_1 B_2, \theta\right)$ for at least one $\mathbf{r}$. If $g\left(\right)$ is injective, then this difference in probability at $\mathbf{r}$ will translate to a difference in probability at the corresponding transformed variable $\hat{A}$. However, because $g\left(\right)$ is a transformation from $\mathbb{R}^N$ to $\mathbb{R}$, it is in most cases not injective. For example, a linear $g\left(\right)$ from $\mathbb{R}^N$ to $\mathbb{R}$ cannot be injective.

Inferring encoding separability from tests of decoding separability

From the results of the previous section, which are summarized in Figure 4, we can conclude that the observation of a violation of decoding separability in a particular brain region is diagnostic of a corresponding

violation of encoding separability. This is because a violation of decoding separability cannot be produced when encoding separability holds. On the other hand, when decoding separability holds nothing can be concluded about encoding separability, as both encoding separability and failures of encoding separability can lead to decoding separability, depending on features of the decoder. This allows an indirect test of encoding separability, which is useful for cases where directly observing encoding separability is difficult (e.g., when indirect measures of neural activity are used, as in fMRI).

Perceptual separability as a form of decoding separability

The concepts of encoding and decoding separability can be linked back to GRT by assuming that perception of a dimensional value is a form of decoding. That is, the key is to assume that the perceptual representation of a stimulus dimension in GRT (the "perceived identity" in Figure 2) is the outcome of decoding a dimensional value from the activity of many channels distributed across the brain (like those shown in Figure 3). This assumption is not new and has proven useful in applications of signal detection theory in the past (Gold and Shadlen, 2001).

More specifically, assume that the perceived value of dimension A, x, is the result of decoding a dimensional value from the activity of many channels distributed across the brain, as in Equation 11. In other words, the perceived value x is a special case of the decoded variable $\hat{A}$, but estimated by readout neurons with the goal of guiding behavioral responses in a perceptual task. Denote the decoding function used by these readout neurons to obtain the perceived value x as $g_R()$, so that

$$x = g_R(\mathbf{r}); \tag{17}$$

this is just a special case of the more general decoding function shown in Equation 11, but limited only to the decoding schemes that can be implemented by real neurons. We have that

$$p(x|A_iB_j) = p(g_R(\mathbf{r})|A_iB_j, \theta) \tag{18}$$

is the marginal distribution of perceptual effects along x. Under these assumptions, perceptual separability (Equation 2) is a form of decoding separability (Equation 14). As a consequence, from Figure 4 we know that *any failure of perceptual separability documented in the literature should be reflected in a failure of encoding separability, in brain areas providing useful information for perceptual identification of dimensional values.* The exact brain regions that provide information to solve a particular task are usually unknown, but we can assume that they encode such information in a relatively transparent (easily decodable) way. The set of potential candidates can be reduced to areas known to provide useful information for behavioral

performance. Novel methods to identify such areas, which combine information about decoded values in a dimension and behavioral response times, have been recently developed (Grootswagers et al., 2018) and seem very promising. For neuroscientists, this opens the opportunity to link new research on the separability of neural representations with decades of accumulated psychophysical research on perceptual separability (Ashby and Soto, 2015).

In addition, as we have seen in the previous section, the common assumption of a Gaussian distribution of perceptual effects (Ashby and Soto, 2015; Soto et al., 2015) is met when the encoding model has additive Gaussian noise and the decoder is linear (see Equation 13), two assumptions that are common in the literature.

Direct and Indirect Tests of Decoding Separability

Direct tests of decoding separability and perceptual separability

Assume that $\mathbf{r}$ is a vector of neural responses encoding dimension A in the brain. If we had access to direct measurements of $\mathbf{r}$ (e.g., firing rates from single cell recordings or a measure of the activity of a homogeneous neural population), we could use an experimenter-defined decoding function $\hat{A} = g_E(\mathbf{r})$ to estimate dimensional values. Obtaining a large number of decoded values $\hat{A}$ for each stimulus A_iB_j allows one to obtain a kernel density estimate (KDE) of $p\left(\hat{A}|A_iB_j,\theta\right)$, represented by $\hat{p}\left(\hat{A}|A_iB_j,\theta\right)$. Comparison of such KDEs constitutes a direct test of decoding separability (Equation 14).

Because perceptual separability is a form of decoding separability (Equation 18), the same procedure can be used to obtain the first available direct test of perceptual separability, when a number of conditions are met. First, the vector $\mathbf{r}$ should include all neural responses encoding dimension A in the brain. Second, for all values of $\mathbf{r}$, $g_E(\mathbf{r}) = g_R(\mathbf{r})$, so that $g_E(\mathbf{r}) = x$ (each experimentally-decoded value is equal to the perceptual effect). As the vector $\mathbf{r}$ cannot be identified and measured using currently available methods and $g_R(\mathbf{r})$ is unknown, both assumptions appear very difficult to meet.

Indirect tests of decoding separability from neuroimaging data

The relations between encoding separability and decoding separability summarized in Figure 4 hold for any decoder, but it can be shown that using a linear decoder allows for a valid test of decoding separability even when indirect measures of neural activity contaminated with measurement error are used, as is the case with fMRI data.

We have assumed the the channel output r_c represents neural activity in a single neuron or a group of neurons with similar properties (e.g., same tuning). Often we do not have access to such direct recordings;

rather, we obtain indirect measures of neural activity, which are some function of the activity of several different neural channels. Let a_m represent an indirect measure of neural activity, where $m = 1, 2, ...M$ indexes different instances of the same type of measure (e.g., different voxels in an fMRI experiment or electrodes in an EEG experiment). The measures can be represented by a vector $\mathbf{a} = [a_1, a_2, ...a_M]$, which is a function of the activity of all channels in the encoding model:

$$\mathbf{a} = \varphi(\mathbf{r}) + \mathbf{e}, \tag{19}$$

where $\mathbf{e}$ is a random vector representing measurement error:

$$\mathbf{e} \sim \epsilon(\theta_e), \tag{20}$$

ϵ denotes the probability distribution of measurement error, which depends on a set of parameters θ_e. Together, Equations 19 and 20 describe the *measurement model* for $\mathbf{a}$.

In a typical multivariate analysis of neuroimaging data, we decode an estimate of a dimensional value $\hat{A}$ directly from $\mathbf{a}$. We can choose to use a linear decoder for this task, so that

$$\begin{aligned} \hat{A} &= \beta + \mathbf{b}^\intercal \mathbf{a} \\ \hat{A} &= \beta + \mathbf{b}^\intercal (\varphi(\mathbf{r}) + \mathbf{e}) \\ \hat{A} &= \beta + \mathbf{b}^\intercal \varphi(\mathbf{r}) + \mathbf{b}^\intercal \mathbf{e} \end{aligned} \tag{21}$$

We can think of the estimate $\hat{A}$ as the sum of two independent random variables: $\hat{A}_\mathbf{r} = \beta + \mathbf{b}^\intercal \varphi(\mathbf{r})$, which depends exclusively on the distribution of $\mathbf{r}$, and $\hat{A}_e = \mathbf{b}^\intercal \mathbf{e}$, which depends exclusively on the error distribution from Equation 20. The variable $\hat{A}_\mathbf{r}$ depends on $\mathbf{r}$ through a composite function. We can think of this composite function as a decoder: $g(\mathbf{r}) = \beta + \mathbf{b}\varphi(\mathbf{r})$ and use $p\left(\hat{A}_\mathbf{r} | A_i B_j, \theta\right)$ to test for decoding separability. Unfortunately, our measurements are contaminated by the variable $\hat{A}_e$ with distribution $p\left(\hat{A}_e | \theta_e\right)$. Because $\hat{A}$ is the sum of two independent random variables, the distribution of $\hat{A}$ is a convolution of the distribution of each of its components:

$$p\left(\hat{A} | A_i B_j, \theta, \theta_e\right) = p\left(\hat{A}_\mathbf{r} | A_i B_j, \theta\right) * p\left(\hat{A}_e | \theta_e\right), \tag{22}$$

where $*$ denotes the convolution integral.

Thus, KDEs obtained from $\hat{A}$ decoded from neuroimaging data reflect the target decoding distribution

convolved with an error distribution. This means that obtaining direct estimates of GRT perceptual distributions from neuroimaging data may not be possible. Still, it is possible to obtain a valid measure of violations of decoding separability.

Without loss of generality, suppose that we want to measure differences between the distributions $p\left(\hat{A}_{\mathbf{r}}|A_1B_1,\theta\right)$ and $p\left(\hat{A}_{\mathbf{r}}|A_1B_2,\theta\right)$. A number of measures of the distance between two probability densities (such as the $L1$, $L2$ and $L\infty$ distances, see Martinez-Camblor & de Una-Alvarez, 2009) start by computing a difference function:

$$\delta\left(\hat{A}_{\mathbf{r}}\right) = p\left(\hat{A}_{\mathbf{r}}|A_1B_1,\theta\right) - p\left(\hat{A}_{\mathbf{r}}|A_1B_2,\theta\right). \tag{23}$$

From neuroimaging data, we obtain estimates of the distributions $p\left(\hat{A}_{\mathbf{r}}|A_1B_1,\theta\right)*p\left(\hat{A}_e|\theta_e\right)$ and $p\left(\hat{A}_{\mathbf{r}}|A_1B_2,\theta\right)*p\left(\hat{A}_e|\theta_e\right)$, where we have assumed that the measurement error model does not change with the value of the stimulus in dimension B. The difference function between these two distributions is:

$$\begin{aligned}
\delta\left(\hat{A}\right) &= p\left(\hat{A}_{\mathbf{r}}|A_1B_1,\theta\right) * p\left(\hat{A}_e|\theta_e\right) \\
&\quad -p\left(\hat{A}_{\mathbf{r}}|A_1B_2,\theta\right) * p\left(\hat{A}_e|\theta_e\right) \\
&= \left[p\left(\hat{A}_{\mathbf{r}}|A_1B_1,\theta\right) - p\left(\hat{A}_{\mathbf{r}}|A_1B_2,\theta\right)\right] * p\left(\hat{A}_e|\theta_e\right) \\
&= \delta\left(\hat{A}_{\mathbf{r}}\right) * p\left(\hat{A}_e|\theta_e\right).
\end{aligned} \tag{24}$$

Thus, the difference between noisy KDEs $\hat{\delta}\left(\hat{A}\right) \approx \delta\left(\hat{A}\right)$ is an estimate of the target difference function $\delta\left(\hat{A}_{\mathbf{r}}\right)$ convolved with the error kernel $p\left(\hat{A}_e|\theta_e\right)$. Note first that if decoding separability holds, then $\delta\left(\hat{A}_{\mathbf{r}}\right) = 0$ and we expect $\hat{\delta}\left(\hat{A}\right)$ to approximate zero for all values of $\hat{A}$ as sample size increases. Any deviations from a constant zero function indicate violations of decoding separability. If decoding separability does not hold and $\delta\left(\hat{A}_{\mathbf{r}}\right) \neq 0$ for some $\hat{A}_{\mathbf{r}}$, then the shape of the error kernel determines how it affects $\delta\left(\hat{A}_{\mathbf{r}}\right)$. Under the common assumption that measurement error $\mathbf{e}$ is Gaussian with zero mean and covariance matrix $\mathbf{\Sigma}_e$, $p\left(\hat{A}_e|\theta_e\right)$ will also be Gaussian with zero mean and variance $\mathbf{b}^\intercal\mathbf{\Sigma}_e\mathbf{b}$. In this case, the convolution attenuates high-frequency fluctuations in the difference function $\delta\left(\hat{A}_{\mathbf{r}}\right)$. In general, the difference $\hat{\delta}\left(\hat{A}\right)$ will capture some deviations from decoding separability, but not necessarily all of them.

In sum, as the number of data points used to obtain KDEs increases, a distance measure based on the function $\hat{\delta}\left(\hat{A}\right)$ will be approximately zero when there are no violations of decoding separability, and any non-zero value will be the consequence of a violation of decoding separability. This makes such a measure a valid indicator of violations of decoding separability. One measure based on $\hat{\delta}\left(\hat{A}\right)$ is the L1 norm:

$$L1 = \int \left| \hat{\delta}\left(\hat{A}\right) \right|, \tag{25}$$

which is the basis for the DDS statistic that we use in our test (see Materials and Methods section).

Linear decoders, which are necessary to obtain a valid indirect test of decoding separability, are also the most widely used in the MVPA literature (Pereira et al., 2009). This allows us to link our framework to this line of research in neuroimaging.

Relation to previous operational definitions of neural independence

Neural representation orthogonality

Suppose that an experimenter believes that two dimensions, A and B, are encoded as two directions in some Q-dimensional space. This space could represent Q measures of neural activity, some transformation of such measures that is of interest (e.g. obtained through multidimensional scaling, or MDS), or some physical coordinate system defined within the brain (e.g., to study topographic maps). Let $\mathbf{h}_A$ be an estimate of the direction along which dimension A is encoded, and $\mathbf{h}_B$ an estimate of the direction along which dimension B is encoded. Several authors have operationalized independence of dimensional encoding as orthogonality of the vectors $\mathbf{h}_A$ and $\mathbf{h}_B$:

$$\mathbf{h}_A \perp \mathbf{h}_B. \tag{26}$$

We call this *neural representation orthogonality*. Note that Equation 26 holds if the Pearson correlation between $\mathbf{h}_A$ and $\mathbf{h}_B$ is zero, as the Pearson correlation of two vectors equals the cosine of their angle. Previous researchers have used both the angle between vectors (Baumann et al., 2011; Kayaert et al., 2005) and their correlation (Hadj-Bouziane et al., 2008) as measures of orthogonality.

Note that we have defined neural representation orthogonality in a very general way, to capture the multiple ways in which the test has been used in the previous literature. In particular, we assume that the vectors $\mathbf{h}_A$ and $\mathbf{h}_B$ may be found through a number of criteria, as long as they represent directions along which dimensions A and B are assumed to be encoded. For example, Baumann et al. (2011) estimated $\mathbf{h}_A$ and $\mathbf{h}_B$ as directions in the physical space of a brain region (the inferior colliculus) along which two dimensions of sound were encoded. Hadj-Bouziane et al. (2008) estimated $\mathbf{h}_A$ and $\mathbf{h}_B$ as the result of unidimensional contrasts (faces $>$ objects and expressive face $>$ neutral face), and thus they represent vectors connecting the average activity for one side of the contrast (e.g., faces) and the other side of the contrast (e.g., objects).

Kayaert et al. (2005) submitted patterns of neural firing rates to MDS, and estimated $\mathbf{h}_A$ and $\mathbf{h}_B$ as directions in the MDS space representing changes in the stimulus dimensions of interest. Another possibility is to estimate $\mathbf{h}_A$ and $\mathbf{h}_B$ as directions in the space of activity measures that separate classes best, as seen in Figure 1. Below we explore this last version of the test, and compare its results to those from the decoding separability test presented earlier.

An important feature of neural representation orthogonality tests that should be considered when interpreting them is that they are best suited to provide evidence of violations of orthogonality, rather than evidence of its presence. More specifically, if $\mathbf{h}_A$ and $\mathbf{h}_B$ were random vectors, then one would expect their correlation to be close to zero (i.e., orthogonal vectors), especially for high-dimensional vectors such as those studied by Hadj-Bouziane et al. (2008). Under such circumstances, a finding of orthogonality is expected even from completely random data. In addition, orthogonality corresponds to a single value (zero correlation or 90-degree angle) and therefore evidence of orthogonality requires special statistical tests that can provide evidence for that specific value (e.g., evidence for the null in a Bayes factor test, or a small confidence interval containing the target value). Such tests have not been used in previous tests of orthogonality. Here, we will explore whether it is possible to find *violations* of orthogonality in our data, rather than trying to find evidence *for* orthogonality as in previous studies.

How is neural representation orthogonality related to the framework presented here? As mentioned earlier, neural representation orthogonality is an operational definition of independence, and as such cannot be linked to the extended GRT framework–or any other framework providing *theoretical definitions* of independence–without further assumptions. A series of assumptions links neural representation orthogonality to perceptual independence, as defined in GRT. *Perceptual independence* of components A and B holds in stimulus A_iB_j if and only if the perceptual effects of A and B are statistically independent; that is, if and only if:

$$p\left(x,y|A_iB_j\right)=p\left(x|A_iB_j\right)p\left(y|A_iB_j\right) \tag{27}$$

When stimulus A_iB_j is presented, it can be represented as a new vector in the same Q-dimensional space that contains $\mathbf{h}_A$ and $\mathbf{h}_B$. If we assume that (i) the projection of this vector onto $\mathbf{h}_A$ and $\mathbf{h}_B$ corresponds to the perceived values of dimensions A and B, then neural representation orthogonality is equivalent to *dimensional orthogonality* (Tanner, 1956; Tucker, 1972). Ashby and Townsend (1986) showed that if, in addition, (ii) the trial-by-trial perceptual effects have a multivariate Gaussian distribution, $p\left(x,y|A_iB_j\right)\sim\mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$, and (iii) $\boldsymbol{\Sigma}$ does not depend on the stimulus (i.e., all perceptual distributions have identical

variance-covariance matrices), then dimensional orthogonality and perceptual independence (as defined in Equation 27) are equivalent.

Assumptions (i)-(iii), linking neural representation orthogonality and perceptual independence, seem extremely strong and hard to meet. Assumption (i) is particularly problematic, as there seems to be no way to guarantee that the estimates $\mathbf{h}_A$ and $\mathbf{h}_B$ must correspond to perceived stimulus dimensions. On the other hand, assumptions (ii) and (iii) are common in the psychophysics literature and they may be justifiable.

It seems even more difficult to link neural representation orthogonality to any property of neural encoding. The reason is that in most applications the Q-dimensional space of $\mathbf{h}_A$ and $\mathbf{h}_B$ is some transformation of the N-dimensional space of the encoding model. Neural representation orthogonality is defined as a 90-degree angle between $\mathbf{h}_A$ and $\mathbf{h}_B$, and only rigid transformations (i.e., not even all linear transformations) can preserve this angle. Thus, unless one assumes that $\mathbf{h}_A$ and $\mathbf{h}_B$ are related to corresponding vectors in the neural encoding space by a combination of rotation, scaling and translation, the neural representation orthogonality test does not provide information about a corresponding property of neural encoding. However, the test might provide useful information about other aspects of brain representation different from encoding, as is the case when the Q-dimensional space of $\mathbf{h}_A$ and $\mathbf{h}_B$ is itself interesting (e.g., physical space in studies of functional brain topography, see Baumann et al., 2011).

In sum, neural representation orthogonality is an operational test of independence of neural representations, and several researchers have used some version of it in past studies. If some strong assumptions are met, the test can be related to the concept of perceptual independence from GRT, which is conceptually distinct from the several forms of separability on which we have focused here. In particular, perceptual independence is a property of a single stimulus. It holds if different stimulus components are processed independently of each other. In contrast, separability is a property of an ensemble of stimuli. It holds if processing of one component is unaffected by changes in other components. In practical terms, we would expect that violations of neural representation orthogonality would be unrelated to violations of decoding and encoding separability, as they measure completely different concepts.

Classification accuracy invariance and generalization

A second operational test of independence of neural representations, more closely related to the separability measures investigated in this article, has been recently used in research on invariance of face representation. In this test, activity patterns in a given brain region are classified according to some target dimension. For example, activity patterns in visual cortex could be classified according to the identity of faces presented during an experiment . Then the classifier is tested with new patterns, produced during presentations of the same identities but with changes in some irrelevant face dimension, such as viewpoint or expression (Anzellotti

and Caramazza, 2014, 2016; Anzellotti et al., 2014). If the classifier's accuracy is significantly above chance, it is concluded that the representations of the target dimension (face identity) are invariant to changes in the irrelevant dimension. A simpler version of the test simply checks for significant classification accuracy using all data (Nestor et al., 2011), but this is much less informative than a test based on generalization after changes in the irrelevant dimension (Anzellotti and Caramazza, 2014).

Formally, let ℓ_i represent a label returned by the classifier indicating that it has estimated that level i of dimension A has been presented, and suppose the experiment includes a total of L_A different levels of dimension A. Then *classification accuracy generalization* is defined in the following way:

$$\begin{aligned} \text{if} \quad & p\left(\ell_i | A_i B_1\right) > \frac{1}{L_A} \\ \text{then} \quad & p\left(\ell_i | A_i B_j\right) > \frac{1}{L_A}, \end{aligned} \tag{28}$$

for all i and j. That is, if the probability of correct classification of the level of A is higher than chance ($\frac{1}{L_A}$) at level 1 of dimension B, then it must be higher than chance at all levels of B for classification accuracy generalization to hold.

It is possible to indirectly relate classification accuracy generalization to encoding separability, through its clear relation to decoding separability. We do this first for the case in which there is access to direct measures of neural activity that are used to estimate $\hat{A}$, as in Equation 11. Because $\hat{A}$ is a noisy estimate that can assume any value in $\mathbb{R}$, a classifier partitions this space into L_A regions, one for each of the values of dimension A included in the experiment. Let $\mathcal{R}_i$ represent the region associated with label ℓ_i, so that the classifier assigns this label to a neural pattern when $\hat{A} \in \mathcal{R}_i$. Each $\mathcal{R}_i$ may be a single continuous interval in $\mathbb{R}$ or composed of several such intervals, and the union of all $\mathcal{R}_i$ completely covers the real line $\mathbb{R}$. When the decoding distribution is known, classification accuracy for level i of dimension A is:

$$P\left(\ell_i | A_i B_j\right) = \int_{\mathcal{R}_i} p\left(\hat{A} | A_i B_j, \theta\right) d\hat{A}. \tag{29}$$

Equation 29 relates classification accuracy to the distribution of decoded values on the target dimension A. From this we know that *if decoding separability holds, then classification accuracy generalization must hold*. This is true because when decoding separability holds, the distribution $p\left(\hat{A} | A_i B_j, \theta\right)$ inside the integral in Equation 29 is the same for all values of j, $P\left(\ell_i | A_i B_j\right)$ is therefore the same for all values of j

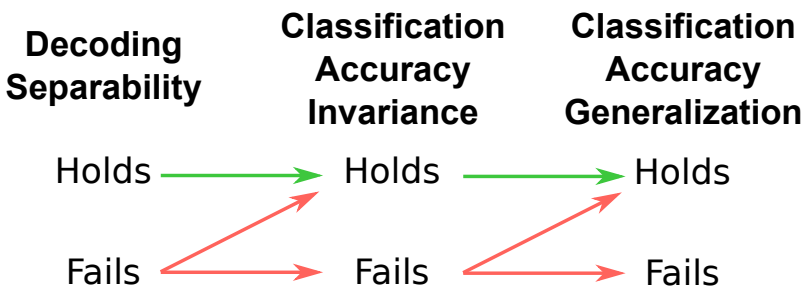

Figure 5: Summary of the relation between decoding separability, classification accuracy invariance and classification accuracy generalization, according to our extension to GRT. Arrows should be interpreted as conditional statements of the form "if X, then Y".

and the relation in Equation 28 holds. On the other hand, *if decoding separability fails, then classification generalization may hold or fail*. Without loss of generality, assume that decoding separability fails because $p\left(\hat{A}|A_iB_1,\theta\right) \neq p\left(\hat{A}|A_iB_2,\theta\right)$. Regardless of the shape of $p\left(\hat{A}|A_iB_1,\theta\right)$ and the region covered by $\mathcal{R}_i$, there are an infinite number of other shapes for $p\left(\hat{A}|A_iB_2,\theta\right)$ that will preserve the area under the curve inside region $\mathcal{R}_i$ constant, making the value of $P\left(\ell_i|A_iB_j\right)$ constant across changes in j, which ensures that the relation in Equation 28 holds. Alternatively, regardless of the shape of $p\left(\hat{A}|A_iB_1,\theta\right)$ and the region covered by $\mathcal{R}_i$, there are also an infinite number of other shapes for $p\left(\hat{A}|A_iB_2,\theta\right)$ that will change the area under the curve inside region $\mathcal{R}_i$, making the value of $P\left(\ell_i|A_iB_j\right)$ change across changes in j. Under such circumstances, the relation in Equation 28 may or may not hold, depending on whether or not the shape of $p\left(\hat{A}|A_iB_2,\theta\right)$ produces an area under the curve inside $\mathcal{R}_i$ that is larger than $\frac{1}{L_A}$.

These considerations point toward an intermediate kind of invariance between decoding separability and classification accuracy generalization, which we call *classification accuracy invariance*, defined as the case in which classification accuracy for levels of dimension A is invariant across changes in the stimulus on a second, irrelevant dimension. That is, classification accuracy invariance of dimension A with respect to dimension B holds if and only if, for all values of i and j:

$$P\left(\ell_i|A_iB_1\right) = P\left(\ell_i|A_iB_2\right) ... = P\left(\ell_i|A_iB_{L_B}\right). \tag{30}$$

Decoding separability (Equation 14), classification accuracy invariance (Equation 30), and classification accuracy generalization (Equation 28) are related to one another as described in Figure 5. The proofs offered earlier relating decoding separability and classification accuracy generalization already include classification accuracy invariance as an intermediate form of invariance for which the relations in Figure 5 hold.

What happens when classification accuracy invariance and generalization are evaluated through indirect

measures of neural activity, such as those obtained from neuroimaging? This is the way in which such tests have been most commonly applied (Anzellotti and Caramazza, 2014, 2016; Anzellotti et al., 2014; Nestor et al., 2011). Remember that in this case, the addition of measurement and noise models (Equations 19 and 20) considerably changes the distribution of $\hat{A}$ estimates obtained from a linear decoder, which is the result of convolving a distribution of decoded values and the distribution of measurement error. This is likely to change the specific classification accuracies $P(\ell_i|A_iB_j)$ but it should not change their relations as defined in Equations 28 and 30, under the assumption that the measurement error model does not change with the value of the stimulus on dimension B.

The theoretical results summarized in Figure 5 reveal two issues with the classification accuracy generalization test as it is currently applied in the neuroimaging literature. The first and most important issue is that finding that classification accuracy generalization holds does not provide any information about encoding separability. On the contrary, what provides information about violations of encoding separability is finding a *violation* of classification accuracy generalization. Thus, while this test seems to be valid and useful, it is currently applied and interpreted in the wrong way (Anzellotti and Caramazza, 2014, 2016; Anzellotti et al., 2014; Nestor et al., 2011). It is possible that finding classification accuracy generalization may provide information about other properties of encoding, but such properties are yet to be identified within a formal framework like the one presented here. Another possibility is that classification accuracy generalization could provide information about encoding separability in special circumstances (i.e., for a specific choice of encoding, decoding, measurement and error models), but again such possibilities are yet to be shown. The second issue with the classification accuracy generalization test is that, given the relations shown in Figures 4 and 5, it provides less information about encoding separability than the decoding separability test proposed in the previous section. In Figure 5, each logical step away from decoding separability implies that a number of violations of encoding separability might go undetected, due to the up-diagonal arrow at each step. Thus, the classification accuracy generalization test is likely to be less sensitive to violations of encoding separability than a decoding separability test. If the goal of a study is to learn about encoding separability, then the wiser decision is to focus on a test of decoding separability, rather than on tests of classification accuracy. An aspect of the lack of sensitivity of the classification accuracy generalization test is the fact that it requires accuracies significantly above chance to be applied and thus should always be applied using an optimal classifier. On the other hand, the decoding separability test offers a sensitive measure of deviations from encoding separability regardless of what decoder is used, including situations in which the decoder is not optimal and/or does not achieve significant classification accuracy.

Summary of Theoretical Results

Here we summarize the previous theoretical results, with an emphasis on how they can be applied to the empirical study of perceptual independence by psychologists and neuroscientists. First, because perceptual separability can be considered a form of decoding separability, and due to the relations summarized in Figure 4, any failure of perceptual separability should be reflected in a failure of encoding separability somewhere in the brain. This means that any psychophysical study reporting a failure of perceptual separability provides a hypothesis to be tested by a neuroscientific study: that a corresponding failure of encoding separability should be found, probably in sensory areas thought to encode the target dimension. Second, such neuroscientific studies can be performed using direct measures of neural activity, such as those provided by single-cell recordings or local field potentials, or indirect measures of neural activity contaminated by measurement error, such as those provided by EEG and fMRI. Using traditional linear decoding strategies on indirect measures of neural activity, the decoded dimensional values still offer a basis for a valid test of decoding separability, and any violation of decoding separability found within a given brain region reflects a violation of encoding separability by the neural population in that region. It must be stressed that a failure of encoding separability is a valid inference that can be made from decoding of neuroimaging data, but such data do not provide a basis to make any strong inferences about the presence of encoding separability. A weak inference can be made, based on the lack of evidence for a violation, but this is analogous to accepting the null in a traditional statistical test. A relatively stronger inference of encoding separability could be made on the basis of assumptions about the neuroimaging measurement model, but researchers should clearly identify such assumptions. Our recommendation to researchers is to be cautious about concluding that separability (or "invariance") holds at the neural level from neuroimaging data, or even from decoding of direct measures of neural activity (Hung et al., 2005).

Finally, we have shown that operational tests of independence available in the literature can be formally defined and re-interpreted within the framework presented here. We showed that, when some strong assumptions are met, the neural representation orthogonality test (Baumann et al., 2011; Hadj-Bouziane et al., 2008; Kayaert et al., 2005) is related to the concept of perceptual independence from the traditional GRT, but it is unlikely to be related to a corresponding property of stimulus encoding. On the other hand, the classification accuracy generalization test promoted by Anzellotti and Caramazza (Anzellotti et al., 2014; Anzellotti and Caramazza, 2014, 2016) can lead to valid inferences about encoding separability. However, the way in which the test has been applied might lead to conclusions of invariance or separability that are in general unjustified, unless one is interested in decoding separability only, and not in the separability of underlying brain representations. In addition, the classification accuracy generalization test is likely to provide

less information than our decoding separability test.

An Application to the Study of Encoding Separability of Face Identity and Expression

Information about a number of properties can be extracted from a single face, including identity and emotional expression. The influential model of Bruce and Young (Bruce and Young, 1986) proposed that these two face dimensions are processed independently, motivating a large number of psychophysical studies aimed at testing this hypothesis (Baudouin et al., 2002; Fitousi and Wenger, 2013; Fox and Barton, 2007; Fox et al., 2008; Ganel and Goshen-Gottstein, 2004; Gao and Maurer, 2011; Lander and Butcher, 2015; Pell and Richards, 2013; Schweinberger and Soukup, 1998; Soto et al., 2015; Soto and Wasserman, 2011; Stoesz and Jakobson, 2013; Wang et al., 2013). Neurobiological theories of visual face processing (Haxby et al., 2000; O'Toole et al., 2002) also propose relatively independent processing of face emotion and identity, through anatomically and functionally differentiated pathways. A ventral pathway projecting from the occipital face area (OFA) to the fusiform face area (FFA) would mediate the processing of invariant aspects of faces, such as identity. A dorsal pathway projecting from the OFA to the posterior superior temporal sulcus (pSTS) would mediate the processing of changeable aspects of faces, such as emotional expression. Recent reviews (Duchaine and Yovel, 2015; Bernstein and Yovel, 2015) conclude that the two pathways are indeed relatively separated and functionally differentiated, with the ventral pathway being involved in the representation of face form information–including invariant aspects of face shape such as identity–, and the dorsal pathway involved in the representation of face motion information–including rapidly changeable aspects of faces such as expression. According to this revised framework, both identity and expression information may be encoded in either pathway, but exactly what information about each dimension is encoded would differ between pathways.

The psychophysical and neurobiological lines of research in this area have remained relatively independent across the years, with no attempt to integrate results across levels of analysis despite the similarity of the central questions guiding their research. In addition, both lines have relied largely on operational definitions of independence that, while having face validity, are usually not linked to any theoretical definition. As indicated in the introduction, this approach makes it difficult to interpret contradictory results.

Thus, the study of independence of face identity and expression is a particularly good testing ground for the theory presented here. Our theory can provide a much-needed theoretical integration across levels of analysis and tests, as well as more rigorous definitions of independence and ways to measure it. In addition, we have recently performed a GRT analysis of psychophysical data to study the perceptual separability of

identity and expression (Soto et al., 2015). The results from that study provide specific predictions to be tested in a neuroimaging study, and thus a proof of concept for our framework. Our behavioral results suggested that, for the stimuli used in that study and after accounting for decisional factors, emotional expression was perceptually separable from identity, but identity was not perceptually separable from emotional expression. From these results, our current framework (see Figure 4) predicts that encoding separability of identity from expression must fail somewhere in the areas representing face information, and that our decoding separability test should be able to find those areas. The predictions regarding encoding separability of emotional expression are less straightforward: as there are no violations of perceptual separability in the behavioral data, we may or may not find violations of encoding separability.

Here, we acquired fMRI data from participants while they looked at the same stimuli and completed the same task as in our previous psychophysical study (see Materials and Methods). This was a simple stimulus identification task, which required participants to pay attention to both identity and expression to attain good performance. Performance in the task during scanning session was high, with a mean of 81.67% (SE = 5.18%). Single-trial estimates of stimulus-related activity were used as input to the decoding separability test described earlier. Because we did not have specific hypotheses about the location of areas showing failures of encoding separability, we performed a whole-brain searchlight analysis (Kriegeskorte et al., 2006), to determine which small circular regions (radius of 3 voxels) showed violations of decoding separability, and therefore violations of encoding separability. To spatially localize violations of encoding separability relative to areas in the face network, we found such areas with the help of a standard functional localizer.

The results from this analysis did not reveal any significant violations of decoding separability, either for identity or emotion. Further exploration revealed that our standardized DDS index was consistently below the value of 0.5 that would be expected under the null hypothesis, suggesting that our method of standarization might have produced an index that is too conservative. We reasoned that one solution would be to use the difference in DDS index between the identity and emotion analyses as the main test statistic, to allow one map to serve as control for the other. This provides only indirect evidence of violations of separability, but would solve the problem of our statistic being overly conservative. Figure 6 shows the main results of this analysis, displayed over a flat cortical map. Face-selective areas found through the functional localizer are outlined in the figure. Outlined in green are face-selective areas showing higher activity during the presentation of faces than during the presentation of other objects. Outlined in red are areas showing higher activity during the presentation of emotional faces than during the presentation of neutral faces. The figure also shows clusters of significant violations of decoding separability, depicted in red-yellow for the identity $>$ emotion contrast. A single large cluster (483 2mm voxels) was found to be significant, covering parts of the left STS and superior temporal gyrus (peak location in MNI coordinates: -60, -14, 2). This

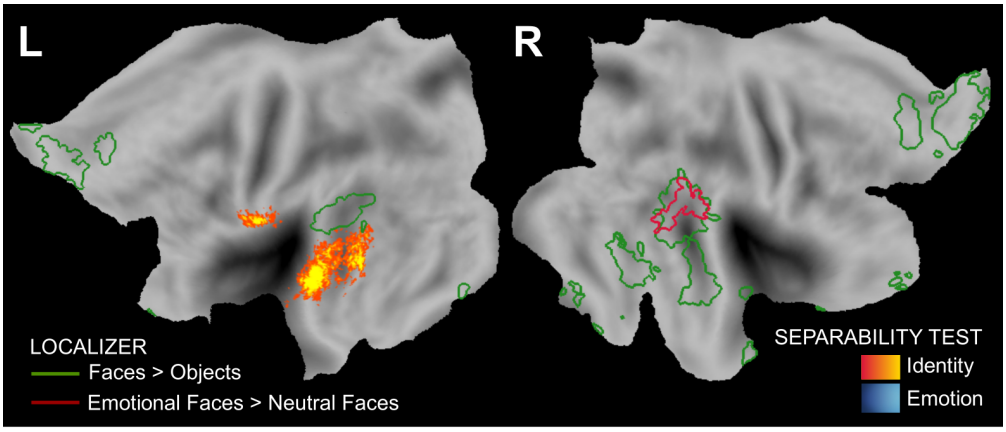

Figure 6: Results of the searchlight decoding separability test. Yellow-red clusters represent regions in which violations of decoding separability of identity were stronger than violations of decoding separability of emotional expression. There were no regions in which violations of decoding separability of emotional expression were stronger than violations of decoding separability of identity. Green and red lines delimit face areas from the functional localizer.

cluster only slightly overlapped with an area of the face network in the pSTS (green contour). No significant violations were found for the emotion > identity contrast.

The results shown in Figure 6 are in line with our predictions, as they provide evidence of stronger violations of decoding separability for identity than for emotional expression, but not the other way around. This asymmetry in the separability of neural representations is analogous to the asymmetry in perceptual separability found in our previous psychophysical study, and thus makes intuitive sense. Although this asymmetry was not a strong prediction from the theory (which simply predicts violations of decoding separability for identity, but is ambiguous about violations of decoding separability for emotional expression), it suggests that there is at least an empirical correspondence between asymmetries of separability in perceptual and brain representations.

Comparison with neural representation orthogonality

We implemented a version of the neural representation orthogonality test (Baumann et al., 2011; Kayaert et al., 2005) discussed earlier. At each searchlight, we measured representation orthogonality by correlating the weights from the classifier used in the previous analyses of identity and emotion. A correlation of zero is equivalent to orthogonality of the two weight vectors, and therefore any deviation from a zero correlation is indicative of a violation of neural representation orthogonality. As mentioned earlier, finding such violations of orthogonality is more informative than finding evidence for orthogonality. The resulting individual orthogonality maps were submitted to the same permutation test previously used for separability maps. No violations of neural representation orthogonality were found in this analysis.

Note that this finding of neural representation orthogonality has been taken by other researchers to mean that information about face identity and emotional expression is represented independently in the visual system (Hadj-Bouziane et al., 2008). Our theoretical results allow us to draw a different conclusion: neural representation orthogonality cannot be easily linked to a corresponding property of stimulus encoding, as even a linear transformation from the space of the encoding model to the space of indirect measures of neural activity does not necessarily preserve angles.

From the point of view of GRT, decoding separability and neural representation orthogonality seem to measure unrelated properties of perception and neural encoding. From the point of view of perception, decoding separability is related to the GRT concept of perceptual separability, whereas neural representation orthogonality is related to the GRT concept of perceptual independence. In both cases, however, the tests are related to the corresponding GRT concepts through a number of strong assumptions. From the point of view of encoding, decoding separability is related to the concept of encoding separability, whereas neural representation orthogonality seems difficult to relate to any property of encoding. For these reasons, we expected the magnitude of violations of orthogonality and violations of separability to be unrelated. To test this hypothesis, we took the group statistical maps obtained from the permutation test in the analysis of decoding separability and computed their Pearson correlation with the corresponding maps from the current analysis of representation orthogonality. There was a small but significant correlation between the orthogonality map and both the map of deviations of separability for identity, r=-0.1162 (p<0.0001), and the map of deviations of separability for emotion, r=0.0666 (p<0.0001). These correlations are significant due to the large number of voxels used to calculate them, but their magnitudes are very small. With these correlations, only 1.35% of the variability in the separability map for identity and 0.44% of the variability in the separability map for expression can be explained by variability in the orthogonality map. Still, the fact that the correlations are significant in real data is important, and some unknown relation between neural representation orthogonality and decoding separability may underlie these results. Future theoretical work will be necessary to clarify these points.

ROI-based decoding separability test

An additional ROI-based analysis was performed, with three goals in mind. First, we wanted to determine whether directly testing face-selective regions would result in some evidence of violations of decoding separability, as the only cluster showing such deviations in the searchlight analysis overlapped very little with face-selective regions from the localizer (see Figure 6). Second, we wanted to more clearly determine whether there are meaningful variations in the amount of separability between different regions. Finally, we wanted to explore the behavior of our decoding separability analysis in control regions. The included ROIs

are face-selective areas (OFA, FFA, STS) and two control regions: V1, which is known to be sensitive to low-level visual features and thus might show deviations of decoding separability (any change in the faces would produce changes in low-level features), and the lateral ventricles, which give us information about the behavior of our statistic when there is very little underlying signal. Some information may be available at the ventricle ROIs that is leaked from adjacent regions, but we would expect that here our statistic should show decoding separability, as the decoding distributions should be almost completely determined by measurement noise.

Results are shown in Figure 7. Figure 7 shows mean standardized DDS values across all ROIs included in the analysis, with error bars representing standard errors of the mean. We tested whether any of these means was significantly higher than the value of 0.5 expected under decoding separability through t-tests (directional, uncorrected). The only ROIs showing significant violations of decoding separability were the left V1 in the analysis of identity, $t(15)=2.04$, $p<.05$, and the right STS in the analysis of emotion, $t(20)=2.05$, $p<.05$. Due to the large number of tests and the fact that our experiment was not originally designed with ROI-based analyses in mind, none of the tests is significant after the application of a correction for multiple comparisons. For this reason, the results shown in Figure 7 should be taken as only suggestive and exploratory. Still, the evidence is encouraging as it suggests that: (1) deviations from decoding separability are not significant in the control areas assumed to include mostly measurement noise (left and right ventricles), (2) deviations from decoding separability are significant in one of the control areas thought to involve such deviations (left V1), and (3) deviations from separability were very low across face-selective areas, with the exception of the right STS, which showed a significant deviation from decoding separability of emotion from identity. Also note that, for most face-selective regions, the mean DDS is consistently below the value of 0.5 which, as mentioned earlier, suggests that the DDS is a conservative measure of failures of decoding separability, at least in areas thought to encode the dimensions under study.

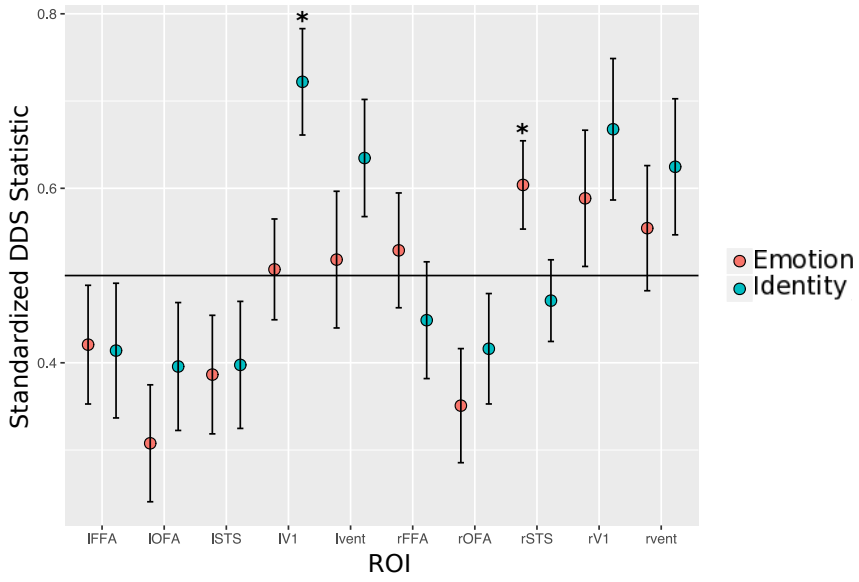

Figure 7: Results of the ROI-based decoding separability test. The y-axis reports the standardized deviations from decoding separability (DDS) statistic. The points represent mean values and the error bars represent standard error of the mean. When decoding separability holds, this index should have a value around 0.5, which is represented with a horizontal dotted line. Mean statistics that were found to be significant (t-test, uncorrected) are marked with an asterisk.

Discussion

Here we have linked multidimensional signal detection theory from psychophysics and encoding models from computational neuroscience within a single theoretical framework. This allowed us, for the first time, to link the results from psychophysical and neurobiological studies aimed at determining independent processing of stimulus properties. Unlike previous approaches, our framework formally specifies the relation between behavioral and neural tests of separability, providing the tools for a truly integrative research approach in the study of independence.

In the past, neuroimaging studies have been limited to a choice between decoding and encoding approaches to data analysis (Naselaris et al., 2011). Decoding approaches focus on answering questions about *what* is encoded in a given brain region, while making no assumptions about *how* exactly that information is encoded; this lack of commitment to an encoding model is both their strength, as they provide useful results regardless of how information is encoded, and their weakness, as they are limited regarding what kind of question they can answer. In contrast, encoding approaches focus on answering questions about *how* a specific stimulus property is encoded in a brain region, but they do this by assuming the encoding model and determining whether it can help to accurately predict data. Their weakness is that there are a very large number of models

that could be tested, and no way of knowing a priori whether the best model is included in the analysis. The theory presented here allowed us to identify some properties of encoding models (i.e., encoding separability) that can be inferred from the results of a decoding study. We hope that future theoretical research in this line will allow researchers to link other properties of encoding models to the results of decoding tests, and more generally to the results of any analysis involving measures that are some transformation of the underlying neural activity, as is the case in fMRI and psychophysics.

Although we focused on developing a decoding separability test, the GRT framework presented here is useful to understand the results of other tests of independence as well (Bate and Bennetts, 2015; Haxby et al., 2000; Ganel et al., 2005; Hadj-Bouziane et al., 2008; Hasselmo et al., 1989; Nestor et al., 2011; Anzellotti et al., 2014; Zhang et al., 2016; Fox et al., 2009; Winston et al., 2004). Here, we re-interpreted two operational tests of independence previously applied in the literature within our extended GRT framework. We showed that, when some strong assumptions are met, the neural representation orthogonality test (Baumann et al., 2011; Hadj-Bouziane et al., 2008; Kayaert et al., 2005) is related to the concept of perceptual independence from the traditional GRT, but it is unlikely to be related to a corresponding property of stimulus encoding. On the other hand, a test based on generalization of classification accuracy (Anzellotti et al., 2014; Anzellotti and Caramazza, 2014, 2016) can provide information about encoding separability. However, the test is likely to provide less information than a decoding separability test and it has been applied incorrectly, yielding conclusions of separability (invariance) that are in general unjustified. Application of our framework to additional operational tests may require the development of models linking neural activity to the specific measurements made in each test.

The framework and test proposed here are applicable not only to fMRI data, but also to the analysis of single-cell recordings, LFPs, EEG and MEG. This breadth of scope across operational definitions and levels of analysis (single neurons, neural populations at many scales, perception, and behavior), which is rarely seen in neuroscience, is a very important contribution of the present work.

We applied our new framework to the study of independent representation of face identity and emotional expression. Previous research found that, for the set of stimuli studied here, identity is not perceptually separable from emotional expression, whereas emotional expression is perceptually separable from identity (Soto et al., 2015). Our results revealed that such lack of perceptual separability is reflected in stronger violations of decoding separability for identity than for emotion in the left temporal cortex, but no stronger violations of decoding separability for emotion than for identity in any brain region.

Several previous fMRI studies have explored the question of whether emotional expression and identity are represented independently in the brain, and an important question is what value is added by a study based on our extended GRT. We believe that our framework provides at least two advantages. The first

advantage is the provision of clear links between the results of neuroimaging and behavioral studies using the same stimuli. No previous study could directly link behavior to neural representation in a meaningful way, as in the present study. We started our study with clear predictions about how fMRI results should reflect the behavioral results, which is preferable to the approach of linking neural and behavioral results through post-hoc theorizing. The second advantage offered by our framework is that it improves our ability to interpret new results in the light of previous results. For example, our data suggest violations of decoding separability of identity from emotion. This does not contradict previous reports of orthogonality of neural representations (Hadj-Bouziane et al., 2008), as we know that decoding separability and neural representation orthogonality measure different concepts. Researchers have also found that emotion can be decoded from areas linked to processing of identity (Nestor et al., 2011; Skerry and Saxe, 2014), and identity can be decoded from areas linked to processing of emotion (Anzellotti and Caramazza, 2017). The issue of whether or not a particular kind of information can be decoded from a brain region is orthogonal to the issue of whether or not it shows encoding separability. Accurate decoding from a particular area indicates that information about a dimension is present in that area but, as indicated earlier, decoding methods are agnostic as to how that information is encoded. The same is true about inaccurate decoding from a particular area. On the other hand, an example of a test that is related to encoding separability is the classification accuracy generalization test of Anzellotti and Caramazza (Anzellotti and Caramazza, 2014; Anzellotti et al., 2014; Anzellotti and Caramazza, 2017). However, this test has not been applied to the study of independence of identity and emotional expression, but rather to the study of identity across changes in viewpoint and modality.

Our application to face perception research is useful to highlight the kinds of questions that can be answered with the new framework and the type of analysis that should be performed to answer such questions. However, there are several limitations of the present study that should be noted. First, we found evidence for our hypothesis through exploratory analyses, as our planned analyses seemed too conservative. More computational research will be necessary to improve our test of decoding separability. Second, results were obtained using a small set of naturalistic stimuli, so they should not be over-generalized. There is no guarantee that the same results will hold for other stimulus sets, and more research is needed before reaching any general conclusion about the separability of identity and emotional expression. Third, our experiment and analyses were performed at the group level. This was done to obtain a statistically-powerful test that is sensitive to violations of separability that are consistent across participants. However, the results may not be representative of individual subjects. We expect that the study of encoding separability at the individual level will require obtaining more data from each participant than what was acquired in the present study.

Our theoretical work might also require further refinement. In particular, the decoding separability test can detect when encoding separability is violated, but it cannot detect when encoding separability holds

(see Figure 4). For many researchers, concluding that a dimension is encoded in a separable manner in a given brain region might be considered more interesting; still, an important contribution of our work is showing that this is not possible through indirect measures of neural activity or psychophysics. Perhaps specific assumptions about the measurement model producing the data will make it possible to establish a more direct link between decoding and encoding separability, but such assumptions need to be clearly spelled out by researchers, and data should be provided to back them up. An additional issue has to do with our proposed DDS statistic, which as indicated above should be studied and improved further.

The notion of independent processing is central to many theories in perceptual and cognitive neuroscience, but its study has lacked the rigor and integration offered by a formal framework, like the one presented here. This framework allows development of theoretically-driven tests of independence of neural representations, which are more clear and rigorous than the operational tests used thus far. The availability of more rigorous definitions and tests to study separability is likely to advance knowledge in a number of areas in visual neuroscience interested in the notions of independence of processing and representation.

References

Amazeen EL, DaSilva F (2005) Psychophysical test for the independence of perception and action. *Journal of Experimental Psychology: Human Perception and Performance* 31:170–182.

Anzellotti S, Caramazza A (2014) The neural mechanisms for the recognition of face identity in humans. *Front. Psychol.* 5:672.

Anzellotti S, Caramazza A (2016) From parts to identity: invariance and sensitivity of face representations to different face halves. *Cereb. Cortex* 26:1900–1909.

Anzellotti S, Caramazza A (2017) Multimodal representations of person identity individuated with fMRI. *Cortex* 89:85–97.

Anzellotti S, Fairhall SL, Caramazza A (2014) Decoding representations of face identity that are tolerant to rotation. *Cereb Cortex* 24:1988–1995.

Ashby FG, Soto FA (2015) Multidimensional signal detection theory In Busemeyer J, Townsend JT, Wang ZJ, Eidels A, editors, *Oxford Handbook of Computational and Mathematical Psychology*, pp. 13–34. Oxford University Press, New York, NY.

Ashby FG, Soto FA (2016) The neural basis of general recognition theory In Houpt JW, Blaha LM, editors, *Mathematical models of perception and cognition, Volume II: A festschrift for James T. Townsend*, pp. 1–31. Routledge, New York, NY.

Ashby FG, Townsend JT (1986) Varieties of perceptual independence. *Psychol. Rev.* 93:154–179.

Bae GY, Flombaum JI (2013) Two items remembered as precisely as one: How integral features can improve visual working memory. *Psychological Science* 24:2038–2047.

Banks WP (2000) Recognition and source memory as multivariate decision processes. *Psychological Science* 11:267–273.

Bate S, Bennetts R (2015) The independence of expression and identity in face-processing: evidence from neuropsychological case studies. *Front. Psychol.* 6:770.

Baudouin JY, Martin F, Tiberghien G, Verlut I, Franck N (2002) Selective attention to facial emotion and identity in schizophrenia. *Neuropsychologia* 40:503–511.

Baumann S, Griffiths TD, Sun L, Petkov CI, Thiele A, Rees A (2011) Orthogonal representation of sound dimensions in the primate midbrain. *Nature Neuroscience* 14:423.

Bernstein M, Yovel G (2015) Two neural pathways of face processing: A critical evaluation of current models. *Neurosci Biobehav Rev* 55:536–546.

Brouwer GJ, Heeger DJ (2009) Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* 29:13992–14003.

Bruce V, Young A (1986) Understanding face recognition. *British Journal of Psychology* 77:305–327.

Cohen DJ (1997) Visual detection and perceptual independence: Assessing color and form. *Attention, Perception, & Psychophysics* 59:623–635.

DeCarlo LT (2003) Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology* 47:292–303.

Demeyer M, Zaenen P, Wagemans J (2007) Low-level correlations between object properties and viewpoint can cause viewpoint-dependent object recognition. *Spatial Vision* 20:79–106.

Duchaine B, Yovel G (2015) A revised neural rramework for face processing. *Annual Review of Vision Science* 1:393–416.

Farris C, Viken RJ, Treat TA (2010) Perceived association between diagnostic and non-diagnostic cues of women's sexual interest: General Recognition Theory predictors of risk for sexual coercion. *Journal of Mathematical Psychology* 54:137–149.

Fitousi D, Wenger MJ (2013) Variants of independence in the perception of facial identity and expression. *Journal of Experimental Psychology: Human Perception and Performance* 39:133–155.

Fox CJ, Barton JJS (2007) What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research* 1127:80–89.

Fox CJ, Iaria G, Barton JJS (2009) Defining the face processing network: optimization of the functional localizer in fMRI. *Hum. Brain Mapp.* 30:1637–1651.

Fox CJ, Oruç I, Barton JJS (2008) It doesn't matter how you feel. The facial identity aftereffect is invariant to changes in facial expression. *Journal of Vision* 8:1–13.

Fox CJ, Moon SY, Iaria G, Barton JJ (2009) The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *NeuroImage* 44:569–580.

Ganel T, Goshen-Gottstein Y (2004) Effects of familiarity on the perceptual integrality of the identity and expression of faces: The parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance* 30:583–596.

Ganel T, Valyear KF, Goshen-Gottstein Y, Goodale MA (2005) The involvement of the "fusiform face area" in processing facial expression. *Neuropsychologia* 43:1645–1654.

Gao X, Maurer D (2011) A comparison of spatial frequency tuning for the recognition of facial identity and facial expressions in adults and children. *Vision Research* 51:508–519.

Garner WR (1974) *The processing of information and structure* Lawrence Erlbaum Associates, New York.

Giordano BL, Visell Y, Yao HY, Hayward V, Cooperstock JR, McAdams S (2012) Identification of walked-upon materials in auditory, kinesthetic, haptic, and audio-haptic conditions a. *The Journal of the Acoustical Society of America* 131:4002–4012.

Gold JI, Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5:10–16.

Goldstone RL (1994) Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General* 123:178–200.

Grootswagers T, Cichy RM, Carlson T (2018) Finding decodable information that is read out in behaviour. *bioRxiv* p. 248583.

Hadj-Bouziane F, Bell AH, Knusten TA, Ungerleider LG, Tootell RBH (2008) Perception of emotional expressions is independent of face selectivity in monkey inferior temporal cortex. *PNAS* 105:5591–5596.

Hasselmo ME, Rolls ET, Baylis GC (1989) The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research* 32:203–218.

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. *Trends in Cognitive Sciences* 4:223–232.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.

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

Kayaert G, Biederman I, Op de Beeck HP (2005) Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience* 22:212–224.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *PNAS* 103:3863–3868.

Lander K, Butcher N (2015) Independence of face identity and expression processing: exploring the role of motion. *Front. Psychol.* 6:255.

Ma WJ (2010) Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research* 50:2308–2319.

Martinez-Camblor P, de Uña-Alvarez J (2009) Non-parametric k-sample tests: Density functions vs distribution functions. *Computational Statistics & Data Analysis* 53:3344–3357.

Mestry N, Wenger MJ, Donnelly N (2012) Identifying sources of configurality in three face processing tasks. *Front. Psychology* 3:456.

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. *Neuroimage* 56:400–410.

Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences* 108:9998–10003.

O'Toole AJ, Roark DA, Abdi H (2002) Recognizing moving faces: A psychological and neural synthesis. *Trends Cogn Sci* 6:261–266.

Pell PJ, Richards A (2013) Overlapping facial expression representations are identity-dependent. *Vision Research* 79:1–7.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage* 45:S199–S209.

Pouget A, Dayan P, Zemel R (2000) Information processing with population codes. *Nature Reviews Neuroscience* 1:125–132.

Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. *Annual Review of Neuroscience* 26:381–410.

Quiroga RQ, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci* 10:173–185.

Richler JJ, Gauthier I, Wenger MJ, Palmeri TJ (2008) Holistic processing of faces: Perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34:328–342.

Rotello CM, Macmillan NA, Reeder JA (2004) Sum-difference theory of remembering and knowing: a two-dimensional signal-detection model. *Psychological Review* 111:588–616.

Rust NC, Stocker AA (2010) Ambiguity and invariance: Two fundamental challenges for visual processing. *Curr Opin Neurobiol* 20:382–388.

Schweinberger SR, Soukup GR (1998) Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception and Performance* 24:1748–1765.

Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. *PNAS* 90:10749–10753.

Silbert NH (2012) Syllable structure and integration of voicing and manner of articulation information in labial consonant identification. *The Journal of the Acoustical Society of America* 131:4076–4086.

Skerry AE, Saxe R (2014) A common neural code for perceived and inferred emotion. *Journal of Neuroscience* 34:15997–16008.

Soto FA, Gershman SJ, Niv Y (2014) Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review* 121:526–558.

Soto FA, Vucovich L, Musgrave R, Ashby FG (2015) General recognition theory with individual differences: A new method for examining perceptual and decisional interactions with an application to face perception. *Psychon Bull Rev* 22:88–111.

Soto FA, Wasserman EA (2011) Asymmetrical interactions in the perception of face identity and emotional expression are not unique to the primate visual system. *Journal of Vision* 11:1–18.

Soto FA, Quintana GR, Pérez-Acosta AM, Ponce FP, Vogel EH (2015) Why are some dimensions integral? Testing two hypotheses through causal learning experiments. *Cognition* 143:163–177.

Stankiewicz BJ (2002) Empirical evidence for independent dimensions in the visual representation of three-dimensional shape. *Journal of Experimental Psychology: Human Perception and Performance* 28:913–932.

Stoesz BM, Jakobson LS (2013) A sex difference in interference between identity and expression judgments with static but not dynamic faces. *J Vis* 13:26.

Tanner WP (1956) Theory of recognition. *Journal of the Acoustical Society of America* 28:882–888.

Thomas R (2001) Perceptual interactions of facial dimensions in speeded classification and identification. *Attention, Perception & Psychophysics* 63:625–650.

Tucker LR (1972) Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika* 37:3–27.

Turner BO, Mumford JA, Poldrack RA, Ashby FG (2012) Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage* 62:1429–1438.

Ungerleider LG, Haxby JV (1994) 'What' and 'where' in the human brain. *Curr Opin Neurobiol* 4:157–165.

Wang Y, Fu X, Johnston RA, Yan Z (2013) Discriminability effect on Garner interference: evidence from recognition of facial identity and expression. *Front. Psychol.* 4:943.

Wenger MJ, Ingvalson EM (2002) A decisional component of holistic encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28:872.

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE (2014) Permutation inference for the general linear model. *NeuroImage* 92:381–397.

Winston JS, Henson RNA, Fine-Goulden MR, Dolan RJ (2004) fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology* 92:1830–1839.

Zhang H, Japee S, Nolan R, Chu C, Liu N, Ungerleider LG (2016) Face-selective regions differ in their ability to classify facial expressions. *NeuroImage* 130:77–90.