Leveraging Psychophysics to Infer the Mechanisms of Encoding Change in Vision*

Jason S. Hays and Fabian A. Soto

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

Abstract

The use of population encoding models has come to dominate the study of human vision, serving as a primary tool for making inferences about neural code changes based on indirect measurements. A popular approach in computational neuroimaging is to use such models to obtain estimates of neural population responses via IEM. Recent research suggests that this approach may be prone to identifiability problems, with multiple mechanisms of encoding change producing similar changes in the estimated population responses. Psychophysical data might be able to provide additional constraints to infer the encoding change mechanism underlying some behavior of interest. Here, we used simulation to explore exactly which of a number of changes in neural population codes could be differentiated from observed changes in psychophysical thresholds. Eight mechanisms of encoding change were under study: specific and nonspecific gain, specific and nonspecific tuning, specific suppression, specific suppression plus gain, and inward and outward tuning shifts. We simulated psychophysical thresholds as a function of both external noise (TvN curves) or stimulus value (TvS curves) for a number of variations of each one of the models. With the exception of specific gain and specific tuning, all mechanisms produced qualitatively different patterns of change in the TvN and TvS curves, suggesting that psychophysical studies can be used as a complement to IEM, and provide constraints on inferences based on the latter. We use our results to provide recommendations for researchers and to re-interpret previous psychophysical data in terms of mechanisms of encoding change.

Keywords: Neural encoding, neural decoding, signal detection theory, psychophysics, inverted encoding modeling.

Introduction

The use of population encoding models has come to dominate the study of human visual perception and neuroscience, serving as a primary tool for making inferences about neural code changes based on indirect measurements (for a review, see Soto and Ashby, 2023). One of the primary benefits of these models is that they can be applied to understand the neurocomputational mechanisms of perceptual processes when more invasive methods are not easily available, as is the case in most human neuroscience studies.

The standard population encoding model (labeled Encoder in Figure 1a; see Pouget et al. 2000, 2003; Ma 2010a; May and Solomon 2015) consists of a population of neural channels (representing a neuron or a population of neurons with similar selectivity), each characterized by

*Manuscript accepted for publication in *Computational Brain and Behavior*

a tuning function that responds more strongly to stimuli that have features similar to its preferred stimulus (i.e., a specific orientation). When a stimulus is presented to the encoder, the set of channels outputs a population response (i.e., a vector of response rates; see Figure 1a), which is affected by internal noise (error bars in the figure). Information about the stimulus is distributed across neural channels in the population response, so that when it is needed for a behavioral task, a decoder must integrate it and recover it (see Figure 1a, Decoded Stimulus).

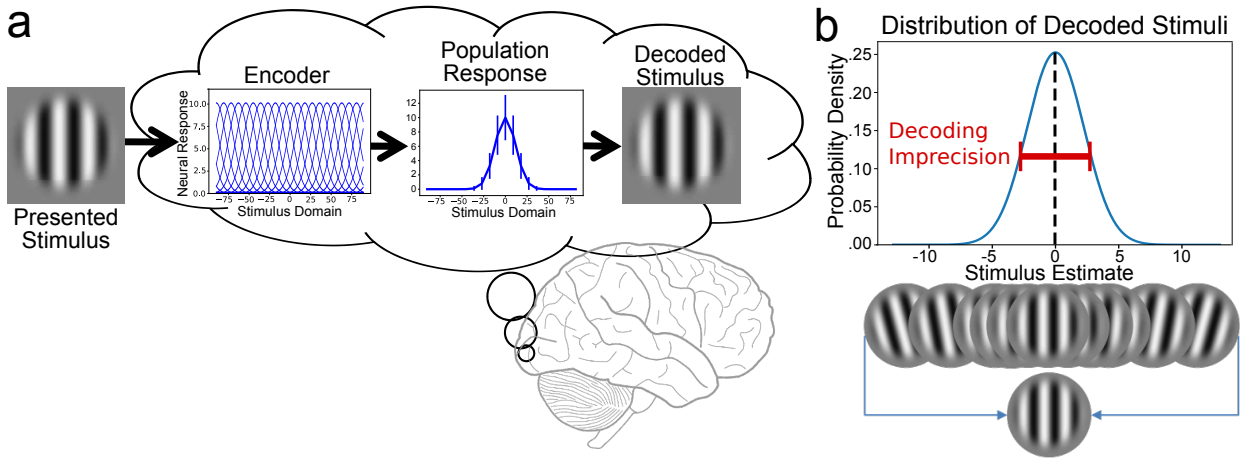

Figure 1: (a) Encoding and decoding of grating orientation. The target stimulus is encoded by a population of neural channels to produce a noisy population response (the error is reported in standard deviations), which is then decoded to produce a single, noisy stimulus estimate. (b) Using optimal decoding, stimulus estimates from many presentations of the same stimulus will be normally distributed around the true stimulus value. The width of the distribution represents decoding imprecision.

A technique commonly used in visual neuroscience to study population codes is IEM (IEM; Brouwer and Heeger, 2009; Sprague et al., 2018, 2019), which is used to make inferences about how neural codes change from a baseline state during and after certain cognitive events (Sprague et al., 2018). The focus of IEM is to obtain an estimate of the population response shown in Figure 1a, rather than directly focusing on properties of the tuning functions in the encoder (Soto and Ashby, 2023). A range of stimuli are presented to participants as they undergo neuroimaging. The same stimuli are presented to an encoding model to obtain average population responses (i.e., without noise). Such responses are then used as predictors for the neuroimaging data using multivariate linear regression and, after inversion of the model, new datasets are used to recover estimates of population responses under multiple conditions (Sprague et al., 2018). For example, IEM has been used to determine how psychological factors such as attention (Garcia et al., 2013; Sprague and Serences, 2013), working memory (Ester et al., 2013), or learning (Byers and Serences, 2014; Ester et al., 2020) influence population responses.

Despite its successes, multiple changes in tuning functions might produce the same change in population response, and thus IEM has known identifiability problems when the goal is to make inferences about changes in neural code (Liu et al., 2018). Some extensions of IEM that have been developed to solve such identifiability problems (Harrison et al., 2023) have had limited success, as they too produce results that are predicted by multiple mechanisms of neural modulation (Wolff and Rademaker, 2024). It could be argued that inferences about neural encoding are exactly what most visual scientists have in mind when they use IEM, as discussion of results usually

focuses on changes in tuning functions rather than on the information available for decoding by downstream neurons provided by the population response (Soto and Ashby, 2023). Given this interest in making inferences about underlying neural codes, what the field requires is new ways to constrain the inferences that can be made through IEM.

From Figure 1 one can see that changes in encoding produce not only changes in population responses, but also in the distribution of decoded stimuli. Given an appropriate choice of decoder, one obtains a type of observer model (Lu and Dosher, 2013) that generates predictions of behavior in psychophysical tasks, which we will call here encoding-decoding observer model (EDOM). There are many possible decoding schemes (e.g., Seung and Sompolinsky, 1993; Pouget et al., 1998; Salinas and Abbott, 1994; Lehky et al., 2013), but optimal decoding via maximum likelihood estimation (MLE; e.g., Deneve et al. 1999; May and Solomon 2015) offers three advantages. First, it is a well-posed statistical problem, and thus this solves the issue of ambiguity regarding whether a change in behavior is due to changes in encoding versus decoding processes. Second, with optimal decoding the distribution of decoded stimuli shown in Figure 1b is a gaussian centered at the true stimulus value, with a standard distribution related to the Fisher information about the stimulus carried by the population code. This links population encoding models with gaussian signal detection theory (SDT), and thus with traditional psychophysical theory. Using SDT, the decoding imprecision shown in Figure 1b is easy to measure, corresponding to the threshold of performance with sensitivity $d' = 1$ measured at the stimulus value. Finally, optimal decoding has shown to be successful in describing human behavior in psychophysical tasks (Dakin et al., 2005; Ling et al., 2009a; Series et al., 2009; Paradiso, 1988).

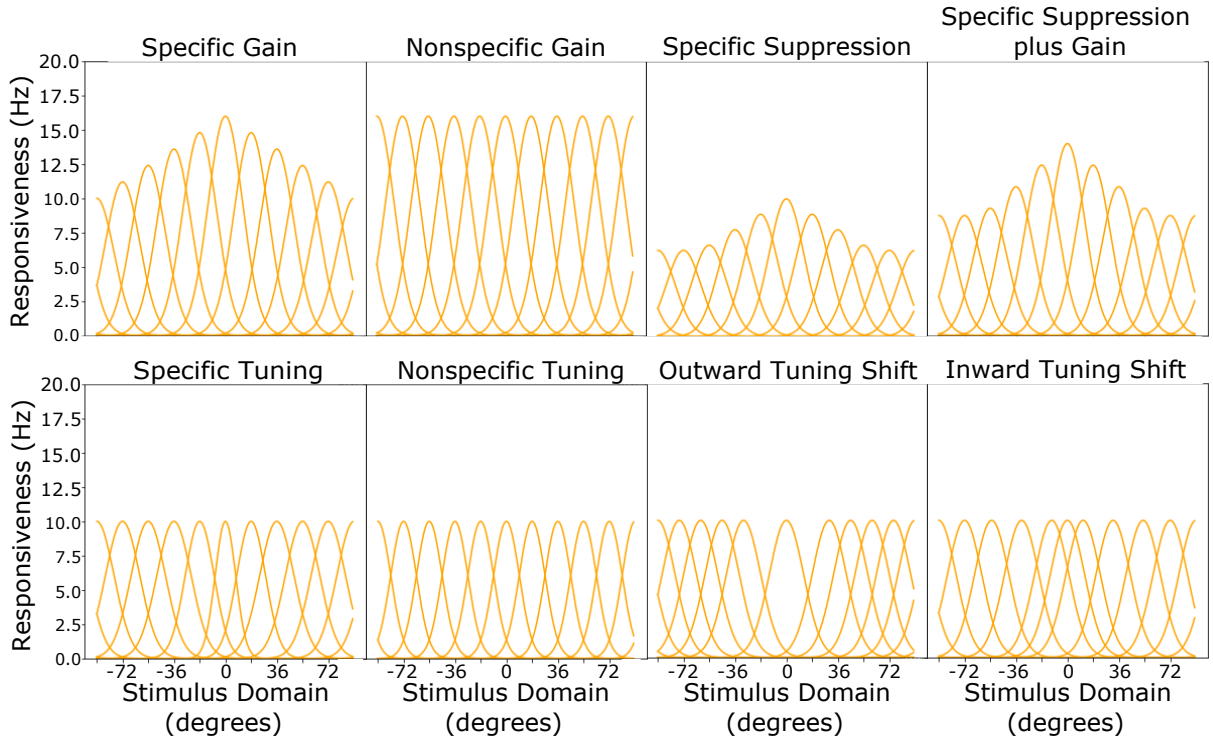

Figure 2: In addition to the homogeneous baseline depicted in the top-left panel of Figure 1a, there were eight different mechanisms of encoding change from baseline under study. Each mechanism is expected to produce unique patterns of thresholds in psychophysical experiments.

Here, we used simulation work with the EDOM to determine which of a number of changes in neural population codes could be differentiated from observed changes in psychophysical thresholds. Figure 2 provides one example for each of the mechanisms of encoding change that we explored. In this and all other figures, we refer to the main stimulus of interest as the target (i.e., we assume that it is specifically targeted by some experimental manipulation, such as training or cueing), which in all cases has a value of zero. In each simulation, the baseline encoding population (i.e., the blue homogeneous population in the encoder of Figure 1a) was modified by a different mechanism of encoding change. Each mechanism affects one of the three parameters of the tuning function (width, height, and center or position) differently. While specific gain, nonspecific gain, and specific suppression all affect the height of tuning functions, specific gain boosts the height for channels with position nearest to the target more than others, specific suppression decreases the height of channels farthest from the target more than others, and nonspecific gain indiscriminately boosts the height of all channels. Specific tuning and nonspecific tuning narrow the width of the tuning functions, but specific tuning narrows the width of channels nearest to the target more than others, while nonspecific tuning indiscriminately narrows the width of all channels. Inward and outward tuning shifts affect the position parameters of the tuning functions; they are "specific" mechanisms insofar as they move each channel's position depending on its distance from the target. Specific suppression plus nonspecific gain implements both mechanisms simultaneously.

These mechanisms of encoding change were selected because they have been proposed in the previous literature to improve task performance in at least some circumstances. Both gain and tuning have been discussed in the attention, conditioning, and category learning literatures (Baldassi and Verghese, 2005; Lawson et al., 2011; Garcia et al., 2013; Serences et al., 2009; Zhang et al., 2010; Stegmann et al., 2019; Yuan et al., 2018; Freedman et al., 2006; Song and Keil, 2014). Specific suppression and specific suppression plus gain have been proposed as external noise reduction mechanisms of attention (Ling et al., 2009b, 2015). Outward tuning shifts have been suggested as mechanisms for improving the performance around category bounds during category learning (Ester et al., 2020). Inward tuning shifts have been suggested as a mechanism of associative learning (Weinberger, 2007, 2011) that causes conditioned stimuli to be over-represented.

In addition, we present an analysis of changes in IEM-estimated population responses under all these mechanisms of encoding change, as a way to understand what kind of information IEM provides to make inferences about underlying codes. From our results we provide guidelines on how estimates of population responses could be interpreted together with psychophysical studies in order to make valid inferences about underlying neural codes.

Materials and Methods

The encoding-decoding observer model (EDOM)

A standard population encoding model (see Figure 1) was used to formally describe how populations of neurons represent a stimulus through their patterns of activity (Pouget et al., 2003, 2000; Ma, 2010b). The response of each neural channel (i.e., neuron or population of neurons with similar properties; indexed by *c*) was described by a tuning function $f_c(s)$ to represent the mean response to any presented stimulus value (*s*). The shape of this function was a Gaussian curve, where r_c^{max} represents the maximum height of the channel, ω_c represents the the standard deviation of the curve (i.e., width), and s_c represents the channel's preferred stimulus:

$$f_c(s) = r_c^{max} \exp\left(-\frac{1}{2}\left(\frac{s - s_c}{\omega_c}\right)^2\right) \tag{1}$$

Channel responses can be influenced by neural noise, which here we describe using a Poisson distribution, independently for each channel (Pouget et al., 2003; Ma, 2010a):

$$p(r_c|s) = \frac{f_c\left(s\right)^{r_c} \exp\left(-f_c\left(s\right)\right)}{r_c!}. \tag{2}$$

When a stimulus with a given value in the dimension s is presented to the model, the output of the population encoding model is a vector of neural responses $\mathbf{r}$, which implicitly encodes information about that stimulus value. As in previous research (e.g., Dakin et al., 2005; Deneve et al., 1999; Ling et al., 2009a; May and Solomon, 2015; Series et al., 2009; Paradiso, 1988) we assume that such information is recovered through optimal decoding via maximum likelihood estimation (MLE).

Generally speaking, the ML estimate is obtained by finding the value of s that maximizes the probability of the population response ($\mathbf{r}$) given a fixed encoding model:

$$\hat{s} = argmax_s p\left(\mathbf{r}|s, \theta\right)$$

, where θ represents a vector with all the fixed model parameters. When neural noise is independent, $p(\mathbf{r}|\hat{s}, \theta) = \prod_{c=1}^{N} p(r_c|\hat{s}, \theta)$, and the sum of the logarithms can be maximized instead:

$$\hat{s} = argmax_s \sum_{c=1}^{N} \ln p\left(r_c|s, \theta\right) \tag{3}$$

For a model with independent Poisson noise and Gaussian tuning functions,we obtain optimally decoded stimuli by finding the value of s that maximizes the following equation:

$$\begin{aligned} \hat{s} &= argmax_s \sum_{c=1}^{N} r_c \ln\left(f_c\left(s\right)\right) - \sum_{c=1}^{N} f_c\left(s\right) \\ \hat{s} &= \sum_{c=1}^{N} r_c \left(\ln\left(r_c^{max}\right) - \frac{1}{2}\left(\frac{s - s_c}{\omega_c}\right)^2\right) - \sum_{c=1}^{N} r_c^{max} \exp\left(-\frac{1}{2}\left(\frac{s - s_c}{\omega_c}\right)^2\right) \end{aligned} \tag{4}$$

Numerical maximization of equation 4 requires providing a starting value for the optimization algorithm. We started the optimization at *s,* to increase speed and precision, as we know that $\hat{s}$ must be in the vicinity of the true stimulus value s.

A property of MLE is that as the number of neural channels increases, the distribution of the decoded stimulus estimate becomes well-approximated by a Gaussian distribution with a mean of *s* and variance equal to:

$$\sigma_{\hat{s}}^2 = I\left(\hat{s}\right)^{-1}$$

where $I(\hat{s})$ is the Fisher information at the ML estimate. The parameter $\sigma_{\hat{s}}^2$ can be estimated through Monte Carlo simulations (Dakin et al., 2005; Ling et al., 2009a). At each repetition $m = 1, 2, ...M$, the model is presented with a given stimulus and the noisy population response is used to decode an estimate of the stimulus value $\hat{s}_m$. Using all M repetitions, one can obtain an estimate of the variance of the distribution of $\hat{s}$ through the following equation:

$$\sigma_{\hat{s}}^2 = \frac{\sum_{m=1}^{M}\left(\hat{s}_m - s\right)^2}{M} \tag{5}$$

Link to sensitivity thresholds

The EDOM presented in the previous section can be linked to SDT if one assumes that the decision variable in SDT is determined by decoding from a population response. The decoded stimulus value $\hat{s}$ corresponds to SDT's decision variable, which is compared to a decision criterion to produce a decision. Because the MLE estimate $\hat{s}$ follows a Gaussian distribution with mean s and variance $\sigma^2_{\hat{s}}$, one can write the following definition of sensitivity:

$$d' = \frac{\delta_s}{\sqrt{\sigma^2_{\hat{s}}}} \tag{6}$$

, where δ_s refers to a small change in the dimension required to detect a change in the target stimulus s with a sensitivity equal to d'. Note that this difference is small enough that we assume an SDT model with a common noise variance for both stimuli, $\sigma^2_{\hat{s}}$. Re-arranging Equation 6 provides an equation to obtain a sensitivity threshold δ_s associated with a given value of d':

$$\delta_s = \frac{\sqrt{\sigma^2_{\hat{s}}}}{d'} \tag{7}$$

In our simulations, we use $d' = 1$, a case in which thresholds are equivalent to the standard deviation of MLE estimates around s (see Figure 1b).

Mechanisms of Encoding Change Under Study

All our simulations follow a similar pattern: starting with a homogeneous baseline set of encoding channels, we generated several variants of the encoding model, each representing a different mechanism of encoding change, and simulated psychophysical functions involving sensitivity thresholds obtained through Equation 7). In this section, we describe how we obtained all the tested population encoding models, and later discuss the simulation procedures in more detail.

To focus on changes in the psychophysical functions that were reliably produced by each mechanism of encoding change, several versions of each model were simulated, and they differed in the magnitude of the effect on encoding and/or the concentration of the effect around the target stimulus. Three effect magnitudes were simulated for each model, and the parameters used to obtain them are summarized in Table 1 and described in more detail below for each model. In addition, for all mechanisms of encoding change classified as "specific," meaning that their effect was strongest at the target stimulus and gradually decreased with distance from the target, we included three additional variants that differed on the concentration of the effect around the target stimulus. In those cases, the "magnitude" variants were factorially combined with the "concentration" variants, to obtain 9 possible versions of each model.

Code Change	Variant A	Variant B	Variant C
Nonspecific Gain (height)	1.3	1.4	1.6
Specific Gain (height)	1.3	1.4	1.6
Nonspecific Tuning (width)	0.6	0.75	0.9
Specific Tuning (width)	0.3	0.4	0.6
Inward Tuning Shift (center)	-10 (-5)	-15 (-7.5)	-20 (-10)
Outward Tuning Shift (center)	10 (5)	15 (7.5)	20 (10)
Specific Suppression	0.77	0.71	0.63
Specific Suppression plus Nonspecific Gain	0.63, 1.4	0.63, 1.3	0.63, 1.2

Table 1: Parameters used to obtain different effect magnitude for each mechanism of encoding change under study. In addition, every "Specific" variation, including the tuning shifts, had 3 additional variants where the effect falloff was increased (1.0, .667, .333—.333 caused the effect to falloff most quickly, 1.0 caused the effect to falloff at the 'edge' of the circular domain—the farthest points from the target). For tuning shifts, the dense populations had reduced parameters (in parentheses) to prevent channels from crossing/overshooting the target. Specific suppression was implemented as a combination of Specific Gain and Nonspecific Gain.

As shown in Table 1, we studied 8 different mechanisms of encoding change: specific/nonspecific gain, specific/nonspecific tuning, inward/outward tuning shifts, as well as specific suppression with and without nonspecific gain (the latter being the only combination of two basic mechanisms, included because of its importance in the literature; see below). For an explanation of why these specific mechanisms were selected, see the Introduction section.

Mechanisms of encoding change categorized as "specific" (including tuning shifts) had an effect that was strongest at the target stimulus and that gradually decreased with distance from the target in the stimulus dimension. In our simulations, this target stimulus was always at the center of the dimension, labeled with a value of zero. The strength of each specific effect decayed linearly with the distance from the target, starting at the target and ending at 1, 2/3, or 1/3 of the half-range of the stimulus dimension. We call this value the *scaling width*, and it represents the inverse of the concentration of the effect around the target. The closer the scaling width is to 1, the more spread out the effect is towards the end points of the dimension. On the other hand, nonspecific changes apply evenly across the domain.

The baseline homogeneous Population

The baseline for all of the simulations was a homogeneous population of channels with $\omega_c = 12$ and $r_{max} = 12$. We assumed a circular stimulus domain varying in degrees, as with grating orientation, so the channels placement ranged from [-90,90) (note that many other stimulus dimensions are described by such circular domains). There were two versions of the baseline homogeneous population, and all the encoding changes specified below were applied to each model. The sparse baseline had 10 channels, placing the channel centers s_c at -90, -72, -54, -36, -18, 0, 18, 36, 54, and 72. The dense baseline had 20 channels, placing the channel centers s_c at -90, -81, -72, -63, -54, -45, -36, -27, -18, -9, 0, 9, 18, 27, 36, 45, 54, 63, 72, and 81.

Nonspecific gain

If the height of the tuning functions is uniformly altered by a scaling factor across the entire stimulus domain, then the change is referred to as nonspecific gain (see Figure 2). The changes in

height directly correspond to changes in the average maximum responsiveness for the channels (i.e., multiplication by r_{max} for all channels). The scaling factors we tested were all above 1 because increases in responsiveness are expected to improve thresholds; thus, we tested: 1.3, 1.4, and 1.6.

Specific gain

If the height of channels that are closest to a target stimulus are increased more than distal channels, then the change is referred to as specific gain (see Figure 2). At the target, r_{max} was multiplied by the the same scaling factors as used for nonspecific gain (see above), but for non-target stimuli the scaling factor was linearly reduced with distance from the target, as described in the introduction to this section.

Nonspecific tuning

If the width of the tuning functions is uniformly narrowed across the entire stimulus domain, then the change is referred to as nonspecific tuning (see Figure 2). We tested scaling factors of 0.6, 0.75, and 0.9, which simply multiplied the parameter ω_c for all channels.

Specific tuning

If the widths of channels that are closest to a target stimulus are narrowed more than distal channels, then the change is referred to as specific tuning (see Figure 2). At the target, ω_c was scaled by factors of 0.3, 0.4, and 0.6, but for non-target stimuli the scaling factor was linearly reduced with distance from the target, as described in the introduction to this section.

Inward tuning shift

If each channel's preferred stimulus moves toward the target stimulus, then the change is referred to as an inward tuning shift (see Figure 2). The maximum shift of the parameter s_c was -10, -15, and -20 for simulations involving the sparse population, and -5, -7.5, and -10 for simulations involving the dense population, where the "-" sign refers to shift towards the target. The dense populations had reduced parameters to prevent channels from crossing/overshooting the target. With a circular domain, there are no true "nonspecific" tuning shifts, as the movement of channels toward the target must necessarily remove channels in areas away from the target. Thus, we simulated only "specific" versions of tuning shift, in which the magnitude of the shift in s_c was linearly reduced with distance from the target, as described in the introduction to this section. In sum, one parameter in these simulations represented the maximum shift of the parameter s_c, immediately next to the target, and another parameter represented the point in the dimension when the effect of tuning shift completely stopped; between these two locations, the tuning shift dropped linearly with distance from the target. Note that the maximum shift is never applied, as the channel closest to the target was either 18 (sparse model) or 9 (dense model) degrees away from it.

Outward tuning shift

If each channel's preferred stimulus moves away from the target stimulus, then the change is referred to as an outward tuning shift (see Figure 2). The maximum shift of the parameter s_c was +10, +15, and +20 for simulations involving the sparse population, and +5, +7.5, and +10 for

simulations involving the dense population, where the "+" sign refers to shift away from the target. With a circular domain, there are no true "nonspecific" tuning shifts, as the movement of channels away from the target must necessarily concentrate channels in areas away from the target. Thus, we simulated only "specific" versions of tuning shift, in which the magnitude of the shift in s_c was linearly reduced with distance from the target, as described in the introduction to this section. In sum, one parameter in these simulations represented the maximum shift of the parameter s_c, immediately next to the target, and another parameter represented the point in the dimension when the effect of tuning shift completely stopped; between these two locations, the tuning shift dropped linearly with distance from the target. Note that the maximum shift is never applied, as the channel closest to the target was either 18 (sparse model) or 9 (dense model) degrees away from it.

Specific suppression

If the height of tuning functions for channels that are closest to the target stimulus are decreased less than for distal channels, then the change is referred to as specific suppression (see Figure 2). Specific suppression can be thought of as the inverse of specific gain: while the target channel's responsiveness is unaffected, the maximum responses of the remaining channels are reduced as a function from their distance to the target (with a larger distance leading to greater suppression). At the target, r^{max} was left the same, but for channels with non-target preferred stimuli the scaling factor was linearly reduced with distance from the target, as described in the introduction to this section. An additional scaling parameter was necessary to determine the maximum level of suppression in r^{max} (or, conversely, the minimum value toward which r^{max} was linearly reduced). These scaling parameters were obtained by inverting the corresponding values used for the gain simulations (i.e., $1/1.3$, $1/1.4$, $1/1.6$; see rounded values in Table 1).

Specific suppression plus gain

Specific suppression plus gain is the only combination of two mechanisms of encoding change that we implemented in our simulations (see Figure 2). The reason was that it has been specifically proposed in the literature as a way in which selective attention improves stimulus encoding, based both in psychophysical (Ling et al., 2009a) and neurophysiological evidence (Martinez-Trujillo and Treue, 2004), and implemented in the influential feature-similarity gain model of attention (Martinez-Trujillo and Treue, 2005). In addition, it has been shown that this mechanism of encoding change is optimal to increase the information available in a neural population about a stimulus that is the target of attention (Nakahara et al., 2001). To implement this combination, we used the highest level of suppression from the previous simulation (i.e., $1/1.6 = 0.63$) and three different levels of nonspecific gain: 1.4, 1.3, and 1.2.

Simulated Inverted Encoding Modeling (IEM) and Analysis of Population Responses

Underlying IEM is the assumption of a linear measurement model (see Soto and Ashby, 2023) linking population responses $\mathbf{r}$ and a vector of indirect brain activity measurements $\mathbf{a}$. This measurement or "mixing" model assumes that activity in an fMRI voxel or EEG channel is a weighted sum of the responses from neural channels:

$$\mathbf{a} = \mathbf{W}\mathbf{r} \tag{8}$$

Here we adopt this measurement model to simulate neuroimaging data across 100 measurements (e.g., voxels, EEG channels). Additionally, we assume additive gaussian noise in each measurement, with a standard deviation of four, which in previous simulation work has produced decoding accuracy similar to that found in studies from primary visual cortex (Soto and Narasiwodeyar, 2023). All our simulations of IEM used the dense encoding model (i.e., with 20 channels).

Using our baseline homogeneous model and the measurement model in Equation 8, we simulated 1,000 trials in response to each of the 20 preferred stimuli s_c (i.e., the center of the tuning functions), for a total of 20,000 trials. Our goal was not to obtain a simulation with realistic experimental parameters, but to test the ability of IEM to recover information about population responses in a "best-case scenario" (hence the large number of stimuli and trials). We used the matrix of simulated activity patterns $\mathbf{a}$ (with noise) together with the mean population responses $\mathbf{r}$ (noiseless, as is customary in applications of IEM) to solve Equation 8 and obtain an estimate of the weight parameters $\hat{\mathbf{W}}$. We then inverted the model to obtain an equation to estimate an underlying population response from measured activity pattern:

$$\hat{\mathbf{r}}_{test} \quad = \quad \mathbf{a}_{test}\hat{\mathbf{W}}^{\top}\left(\hat{\mathbf{W}}\hat{\mathbf{W}}^{\top}\right)^{-1}, \tag{9}$$

where $\mathbf{a}_{test}$ represents an activity pattern measured during a test trial, independent of trials used to estimate $\hat{\mathbf{W}}$. We simulated 1,000 activity patterns in response to the target stimulus (i.e., a stimulus value of zero) for each of the models under study (including baseline; see previous section), and estimated the underlying population response using Equation 9. The final estimates of population responses were obtained by averaging across all the test activity patterns. Presentation of only the target stimulus is a deviation from standard practice in population encoding modeling, which involves presentation of a range of stimuli covering the whole stimulus domain. The population responses obtained for each stimulus are then shifted, to have the same center, and averaged. This is a sensible approach under the assumption that the mechanism of encoding change is nonspecific, operating equally across all channels. However, when the underlying mechanism is specific, operating differently across channels, a shifted average could recover a population response that does not match any of the underlying population responses to each stimulus. In many cases (e.g., specific gain, specific tuning, specific suppression plus gain), the effect observed in the shifted average is reduced or eliminated compared to the effect observed at the target. For this reason, here we study the effect at the target.

Population responses obtained directly from the model and estimated through IEM were fitted to a function which has the Gaussian as a special case, but that can also produce curves with lower or higher curvature at the peak than the Gaussian (including exponential functions, when the curvature at the peak is very low). The function is defined by the following equation:

$$f_P\left(s\right) = r_P^{min} + r_P^{max}\exp\left(-\left|\frac{s}{\omega_P}\right|^m\right), \tag{10}$$

where r_P^{min} is the baseline level of responding (set to zero, as no observed curves showed higher baseline), r_P^{max} is the peak level of responding, ω_P determines the width of the function (the inverse of its decay slope), m determines the width of the function at the peak (with $m = 2$ representing a Gaussian function, and $m = 1$ representing an exponential decay function), and s is the stimulus value. A similar function has been fitted to estimates of population responses obtained using IEM (Ester et al., 2020; O'Bryan et al., 2024).

The population curve in Equation 10 was fit to the population responses using least squares estimation with the bound-constrained optimization method (L-BFGS-B) included with *optim* in R v. 3.5.1. The lower bounds enforced for the three fitted parameters were as follows: $r_P^{max} = 5$, $\omega_P = 1$, and $m = 1$. The optimization procedure was repeated fifty times, each time with a different starting set of parameters, and the best-fitting parameters across all repetitions were included in further analyses. Starting parameters for the optimization algorithm were obtained by adding random values to the lower bound parameters described above, using a normal distribution with mean equal to zero and variance equal to five. The results from this analysis (Figure 5) were plotted using *ggplot2* v. 3.1.

The curves obtained from IEM estimates were evaluated in the following way: (1) using the best-fitting parameters, the curves were evaluated at 200 evenly-spaced stimulus values; (2) the baseline curve was subtracted from the curves corresponding to each mechanism of encoding change, resulting in a difference curve that captures changes in the population responses observed with a change in encoding; (3) each difference curve was compared to a corresponding curve computed from population responses, using thesquare of the Pearson correlation, r^2, which ranges from 0 to 1 and quantifies how well the changes in IEM population curves capture changes in the true underlying population curves.

Simulated Psychophysical Functions

Threshold vs Noise (TvN) curve In these simulations, we obtained thresholds from each model under study as explained above. At each repetition $m = 1, 2, ...M$, we presented the model with both the target stimulus and a pattern of external noise with a particular magnitude. We assumed that the pattern of external noise had two effects on the response of the model. First, because a pattern of external noise can sometimes approximate each channel's preferred stimulus to a certain extent, we stimulated each channel with its preferred stimulus and then scaled down the population response through a random weight. Each weight was randomly sampled from a normal distribution with mean zero and a standard deviation σ_E, which was varied to produce different levels of external noise. The second effect of the pattern of external noise would be to degrade the presented target stimulus (e.g., see grating stimuli in Figure 7). Because the target cannot be represented perfectly in the presence of external noise, its produced population response was also scaled by a weight equal to $1 - \sigma_E$.

There were 375 levels for the external noise parameter σ_E, ranging from 0.0 (no external noise) to approximately 0.5 (0.513; a point where the noise response vectors had approximately the same strength as the target). The levels of external noise were evenly spaced along a log scale, meaning that values were closer together around 0 and became increasingly more separated as they approached the maximum.

After obtaining thresholds at each of the 375 levels of external noise, B-spline smoothing was applied to the resulting TvN curves, in an attempt to obtain better approximations to an "idealized" TvN curve from each model. For all TvN curves, a smoothing parameter of 50 was selected. The goal behind parameter selection was to represent the ideal (i.e., smooth) curve as closely as possible without disrupting the relative threshold relations between the curves.

The dense population models in general produced lower thresholds than the sparse population models. To aid comparison across models, we transformed the TvN curves obtained using sparse models, so that they would be in the scale shown by curves obtained using dense models. We obtained a scaling vector by dividing the sparse baseline TvN by the dense baseline TvN, after smoothing, at each of the levels of external noise. This scaling vector was then applied to every sparse TvN curve, by applying the Hadamard product between both vectors.

Threshold vs Stimulus (TvS) curve

In these simulations, we presented 181 stimuli, ranging evenly from -90 to 90, to each model under study. No external noise was presented, but neural noise was still present. As before, B-spline smoothing was applied to the obtained TvS curves. Because the stimulus domain is circular, thresholds were appended to both sides to prevent the B-splines from curling at the ends. We used different smoothing parameters for different simulations, again with the goal of representing the ideal (i.e., smooth) curve as closely as possible without disrupting the relative threshold relations between the curves. For the dense populations, the standard smoothing weight was 3, but nonspecific tuning only needed a weight of 2, and specific suppression needed a weight of 4. For the sparse populations, the standard smoothing weight was 6, but nonspecific tuning only needed a weight of 4, specific suppression needed a weight of 11, inward shift needed a weight of 7, and specific gain needed a weight of 50. The goal behind selection of smoothing weights was to present smooth psychophysical curves that would make it easy to see how each curve was affected by a given mechanism of encoding change, without disrupting the relative threshold relations among the curves.

To aid comparison across models, we transformed the TvS curves obtained using sparse models, so that they would be in the scale shown by curves obtained using dense models, in the same way as described for TvN curves above.

Simulation environment specifications

The simulations were run on a Titan W375 Workstation PC with 32 dual-core (64 cores and 128 threads total) AMD EPYC 7551 2.0GHz (3.0GHz Turbo) 64MB Cache processors, running Ubuntu 18.04.4 LTS. Simulations were programmed using Python 3.7.2, extended with numpy v. 1.15.4, scipy v. 1.1.0. For parallelization, ipyparallel v. 6.2.3 was used in conjunction with jupyter-client v. 5.2.4 and notebook v. 5.7.4 integration. The pandas v. 0.23.4 module was used to save, read, and manage data, and matplotlib v. 3.0.2 was used to plot data. Python and all modules were obtained through the Anaconda distribution v. 4.5.12.

Results

IEM-Estimated Population Response Curves

As indicated in the introduction, the main goal of IEM is to obtain estimates of population responses (Sprague et al., 2018). For this reason, we used IEM to estimate such population responses after presentation of the target stimulus, from each one of the models under study. To evaluate the ability of IEM to reproduce the true underlying population responses, we also obtained those directly from each of the encoding models. The obtained responses are shown in Figure 3, with the true responses in the top panel and the IEM-estimated responses at the bottom. It can be seen that the true population responses are in general wider than those recovered by IEM. As a consequence, some aspects of the true population responses are not captured by IEM, such as the drop below baseline observed for all the gain and suppression models (top row of models) and for inward shift. The IEM-estimated responses also show two features not observed in the true population responses. First, a drop below baseline for the peak response in the tuning models, specific suppression, and most notoriously outward shift, and a corresponding increase above baseline for inward shift. This suggests that a change in the peak response estimated from IEM could be due to a number of underlying mechanisms, rather than only gain, as is the case in the

true underlying populations. Second, a “mexican hat” pattern for non-specific tuning, in which responses drop below zero away from the target stimulus, before rising again to a value near zero for stimuli farther away from the target. This pattern is theoretically important, as some researchers associate it with lateral inhibition (e.g., Antov et al., 2020), but in the context of IEM it seems to result purely from features of the estimation procedure.

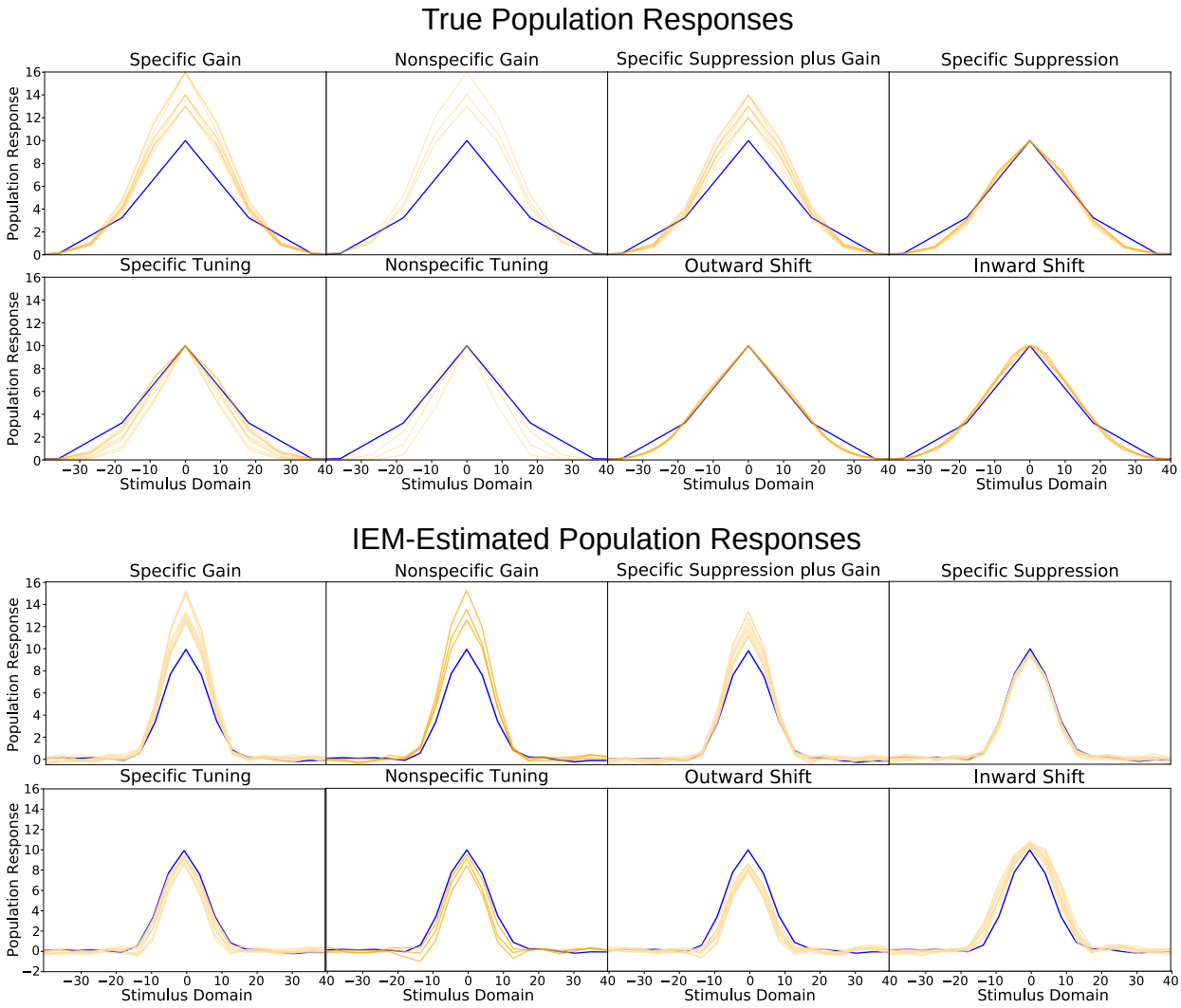

Figure 3: Population responses directly obtained from each of the encoding models under study (top) and estimated through IEM (bottom). Blue curves correspond to baseline models and orange curves to encoding changes implemented in a dense population (20 channels).

Nevertheless, the procedure seems to capture nicely important features of the true population responses. Importantly, it can clearly distinguish between gain mechanisms, which show an increase in response, and tuning mechanisms, which show a narrowing of the curve.

To evaluate how well IEM recovers features of the true population responses, we performed an analysis in which each of the population responses shown in Figure 3 were fitted to a model of the population response curve (see Equation 10), similar to that used in previous IEM (e.g., Ester et al., 2020), in that it allows to obtain separate estimates of the population response’s height and width,

but also provides estimates of the curvature (i.e. width) of the curve at the peak. The resulting curves were evaluated at 200 evenly-spaced stimulus values. The baseline curve was subtracted from the curves corresponding to each mechanism of encoding change (i.e., orange curves in Figure 3), resulting in a difference curve that captures changes in the population responses observed with a change in encoding. Each difference curve was compared to a corresponding curve computed from dense population responses (response of 200 evenly-spaced channels to the target stimulus), using r^2 (i.e., Pearson correlation squared, or coefficient of determination). A focus on difference curves allows us to determine how IEM captures changes from baseline rather than the general shape of the population responses, which is known to be influenced by assumptions in the model (Gardner and Liu, 2019). Figure 4 shows the results of this analysis.

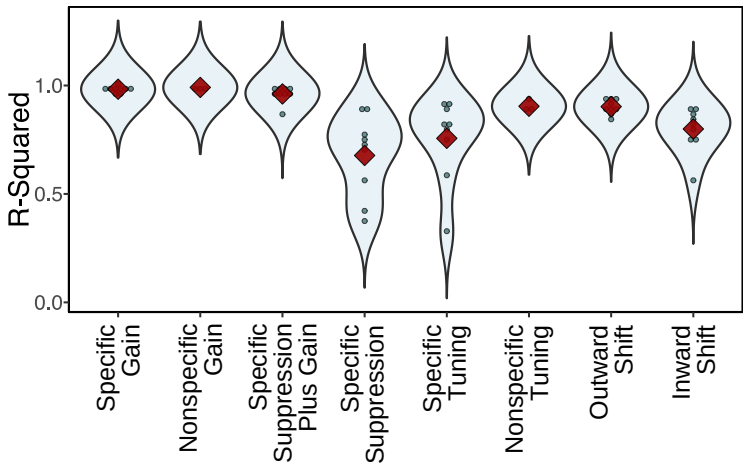

Figure 4: Correspondence of the population responses estimated through IEM and the true underlying population responses, measured through r^2.

The ability of IEM to recover changes in the true population curves was excellent for specific gain (r^2 mean = .98; range: .97-1.0) and non-specific gain (r^2 mean = .99; range: .98-.99), and quite good for specific suppression plus gain (r^2 mean = .96; range: .87-.99), non-specific tuning (r^2 mean = .90; range: .89-.93) and outward shift (r^2 mean = .90; range: .86-.94). Note that these values should be taken as upper bounds for the values that could be obtained through IEM, as our simulations assumed ideal experimental settings hard to be reproduced in real studies. Three model types showed values of r^2 in some simulations that might be considered unsatisfactory: specific suppression (r^2 mean = .68; range: .37-.89), specific tuning (r^2 mean = .76; range: .32-.92), and inward shift (r^2 mean = .8; range: .57-.90). In all these cases, some changes in the model were captured with good accuracy, but others were captured poorly. The poor results are not a result of the accuracy of the model of the population curve used to fit the IEM results, as an additional analysis showed that r^2 values never drop below .96 when the model was fitted to true population responses evaluated at 20 stimulus values (i.e., those shown at the top of Figure 3) and analyzed as with IEM-estimated responses, with most r^2 values being higher than .99. This additional analysis indicates that, in most cases, our model of the population curve provided a good description of population responses obtained from the models and we can interpret the shape of true population responses through the recovered parameters.

Such recovered parameters were subtracted from the corresponding parameters obtained from baseline models, and the distributions of these differences are shown in Figure 5. Results for the

true population responses are shown in the left panel, while results for the IEM-estimated population responses are shown in the right panel. We must first note that many applications of IEM focus simply on whether the population response changes in height versus width in a given experimental condition (e.g., attention). It can be seen from the Figure that an increase in height of the curve was always diagnostic of a gain mechanism at the neural encoding level (specific gain, nonspecific gain, or specific suppression plus gain), and results in the true population response matched those in the IEM-estimated responses. However, IEM clearly shows a bias to produce an increase in height under inward shift that does not occur in the underlying population responses, which means that this possibility should be eliminated in another way before reaching a definitive conclusion. On the other hand, changes in the width of a curve are not very diagnostic of changes at the neural encoding level. In the true population curves, we see narrowing of the curve width due both to changes in tuning (specific and nonspecific) and other mechanisms (specific gain, specific suppression plus gain, specific suppression). These results are only partially mirrored in the IEM-estimated population curves, many of which showed widening of the curve that was not present in the true population curves, either with some parameter settings (specific gain, nonspecific gain, and specific tuning), or all of them (inward shift). IEM consistently and correctly predicted a narrowing of the tuning curve only for nonspecific tuning, but also consistently and incorrectly predicted such narrowing for outward shift. This mirrors the point made by previous simulation work (Liu et al., 2018) showing that changes in tuning of recovered population responses are not diagnostic of changes in tuning of encoding neural channels. IEM did not recover the true curvature at the peak response in most cases, rendering this parameter an invalid indicator of changes in true population responses.

The results shown in Figure 5 do reveal some information about the underlying mechanisms of encoding change, even though they provide that information only indirectly. For example, a gain mechanism always produces an increase above baseline of the curve height parameter, accompanied by small changes in curve width, whereas a tuning mechanism in most cases produces a reduction from baseline of the curve width parameter, accompanied by a drop in curve height. This shows that IEM is well-suited in most cases to distinguish between the two general categories of gain and tuning, which might be of theoretical interest in many applications. However, note that there is no qualitative pattern of parameter changes (increment, decrement, or no change) obtained through IEM that is specific to a particular mechanism of encoding change. Without additional data, IEM does not provide information about what specific mechanism of encoding change underlies changes observed in neuroimaging measurements, which seems to be in line with the recommendation of not making inferences about underlying channel tuning functions from recovered population responses (Sprague et al., 2018). An additional caveat revealed by our simulations is that inferences about true population responses based on IEM might in some cases also be invalid, as shown by the results in Figure 4.

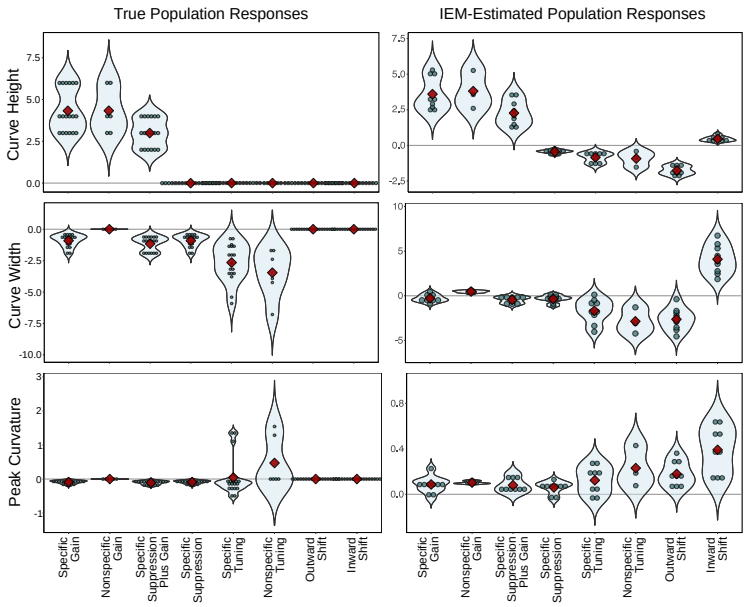

Figure 5: Distributions of parameters recovered from the population response to the target stimulus, presented separately for each model under study. The values represent the difference in estimated parameter between a particular model and its corresponding baseline, with values higher than zero representing "taller" or "wider" curves, and values below zero representing "shorter" or "narrower" curves. Each dot represents a different model variation and the red diamond represents the mean of the distribution.

On the other hand, we would be amiss if we did not point out that the IEM approach has more potential to provide information about encoding changes than what standard practice permits. First, note that the parameters provided in Figure 5 were obtained from population responses to a single target stimulus. In real applications, multiple stimuli with different values in the dimension of interest are presented, the population responses are estimated for each one of them, shifted to have a common zero mean, and averaged to obtain a single estimate of the population response. This practice allows to obtain better estimates of the population response *only when nonspecific mechanisms are involved.* To understand why this is the case, Figure 6 shows the population responses obtained by presenting seven evenly-spaced stimuli, starting from the target and moving towards the right side of the dimension, to an ultra-dense version of the population encoding models used here. This allowed to obtain smooth population responses showing all the information ideally available from an IEM experiment. Note first that when nonspecific gain or tuning are involved, all population responses have the same shape and their average is a good estimate of that curve. For all other cases, population responses vary with presented stimulus. For example, in specific gain population responses drop in height as the stimulus gets away from the target, in specific tuning population responses widen as the stimulus gets away from the target, and so on.

Thus, the common practice of averaging shifted estimates of population responses has three undesirable consequences. First, when a specific mechanism is involved, averaging produces biased estimates of any of the true underlying population responses. Second, averaging may reduce the size of the effect of an experimental factor on population responses, and in some cases it might even artificially get rid of such effect (e.g., when averaging responses higher and lower from baseline, in the case of specific suppression plus gain). Finally, averaging discards an important amount of information about the underlying mechanism of encoding change. For

example, the parameters recovered from a single curve in Figure 5 cannot distinguish between specific gain and specific suppression plus gain, but the latter mechanism is the only one that produces suppressed population responses away from the target in Figure 6. Similarly, specific suppression, specific tuning, and nonspecific tuning cannot be distinguished from the parameters shown in Figure 5, but they clearly produce different patterns of population responses as a function of stimulus value in Figure 6.

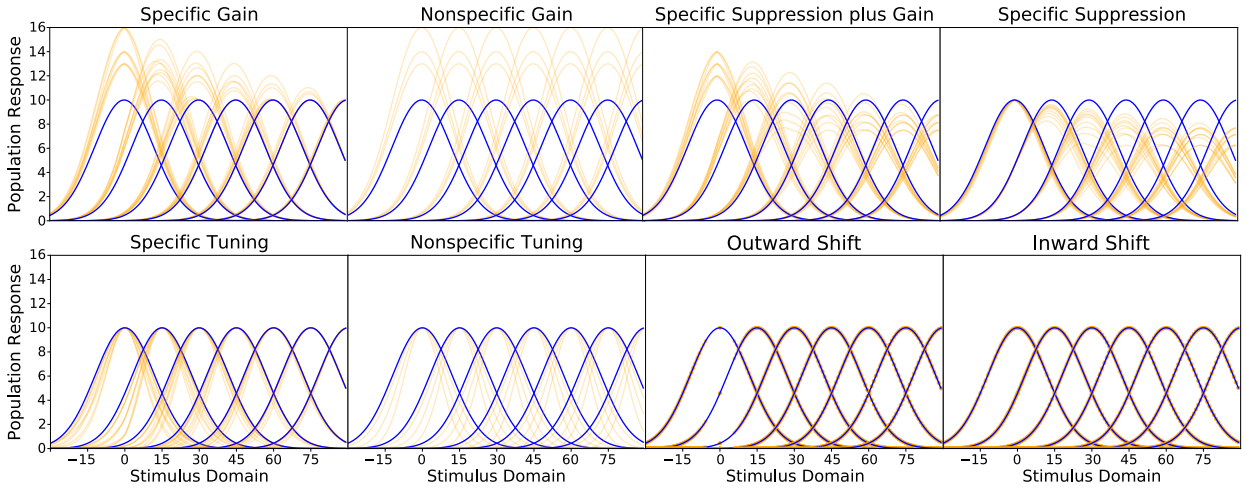

Figure 6: Idealized population responses obtained from ultra-dense population encoding models. They show that the current practice of averaging population responses from different stimuli reduces the information available about underlying mechanisms of encoding change.

Thus, it seems like the inability to differentiate mechanisms of encoding change using IEM may be due in part to the practice of averaging population responses obtained by presentation of different stimuli. A much better approach to differentiate between the different mechanisms in Figure 2 would be to estimate a different population response for a number of stimulus values along the dimension. This, however, would require many times the amount of data that is usually obtained in IEM studies, and might not solve other identifiability issues highlighted by recent research (Liu et al., 2018) and our simulations. In addition, traditional inverted encoding analyses cannot provide information about tuning shift mechanisms, and it is not clear whether special analyses proposed to obtain such information for outward tuning shifts (Ester et al., 2020) would similarly work to infer inward tuning shifts.

Threshold vs Noise (TvN) Curves

The equivalent-noise paradigm (Pelli, 1981) is a widely-used psychophysical tool that involves measuring the sensitivity of the visual system to changes in a given variable (e.g., grating orientation) at many different values of external noise that is added to the stimulus. Sensitivity is usually measured via thresholds, which as indicated earlier are estimates of decoding imprecision. Figure 7 shows an explanation of the resulting Threshold versus external Noise (TvN) curve (also called TvC curve, where C stands for “Contrast Noise”). The typical shape of this TvN curve has a flat section at low levels of external noise, where performance is almost exclusively limited by internal sources of noise, such as neural noise in our simulations. This is followed by a curved section where external noise starts exerting its influence, and ends with a linearly-increasing section

where performance is almost exclusively determined by the level of external noise.

TvN curves have been used in the past to characterize psychophysical observer models (Lu and Dosher, 2013), as well as encoding/decoding observer models like the one presented in the introduction (e.g., Dakin et al., 2005; Ling et al., 2009a). In the context of our study, the TvN curve has been shown to provide useful information about mechanisms of change in encoding populations. The reason for its usefulness is that different changes in encoding should produce different changes in the curve, depending on whether they affect internal or external noise. For example, Ling et al. (2009a) found through simulations that a nonspecific gain mechanism of attention would specifically suppress internal noise, making the TvN curve drop only at low levels of external noise (green area in Figure 7). On the other hand, a "tuning" mechanism of attention (what we have called specific suppression with nonspecific gain) would specifically suppress external noise, making the TvN curve drop only at high levels of external noise.

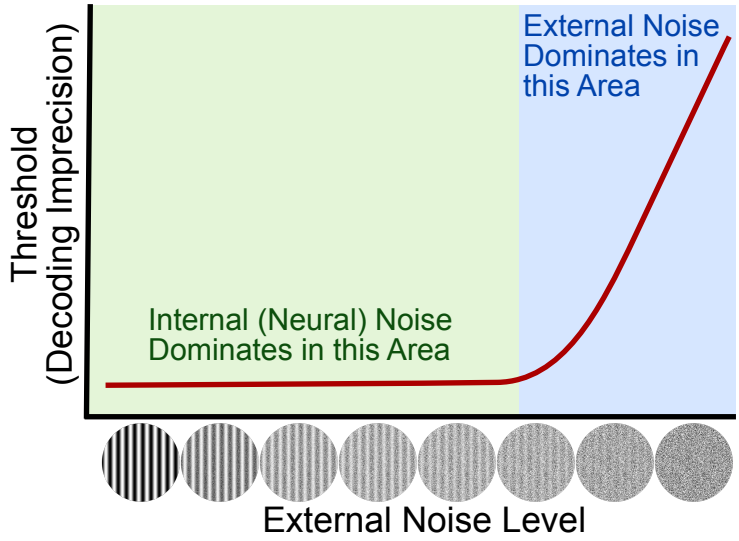

Figure 7: Explanation of a Threshold versus external Noise (TvN) curve.

The goal of the present set of simulations was to expand previous work (Dakin et al., 2005; Ling et al., 2009a) and determine how *all* the changes in population codes shown in Figure 2 affect the shape of the TvN function. The obtained TvN curves are shown in Figure 8, with each panel representing a different mechanism of encoding change. Baseline curves are shown in blue, whereas curves obtained from the models in Figure 2 are shown in either orange (dense population) or green (sparse population). In the rest of this section, we describe the results shown in Figure 8 in detail for each model.

From the point of view of the decoder, information about differences in stimulus values comes from differences in the responses of each channel. The slope of the tuning function at a particular value of the stimulus determines its sensitivity to changes in the stimulus (i.e., how different will be the response as a function of a small change in the stimulus). For a range of values, such slope can increase by increasing the height or decreasing the width of the tuning function.

Broadband noise involves adding random stimulation across all neural channels. Again, the slope of the tuning function determines to what extent small changes in random stimulation produce a large change in the channel's response. In this case, however, a higher slope for channels other than those responding to a target stimulus means more influence of small variations in ran-

dom noise, which decreases the precision of decoding for that target.

In “nonspecific” mechanisms of encoding change, the slope of tuning functions is affected equally across all channels. At low levels of external noise, this improves decoding precision, but at high levels of external noise, the effect of the noise on channels surrounding the target is stronger, worsening decoding precision. On the other hand, in “specific” mechanisms of encoding change, the slope of channels that respond to the target stimulus is affected more than the slope of surrounding channels that respond to noise, producing an overall improvement in performance across all noise levels.

Following is a description of results separately for each specific mechanism of encoding change under study.

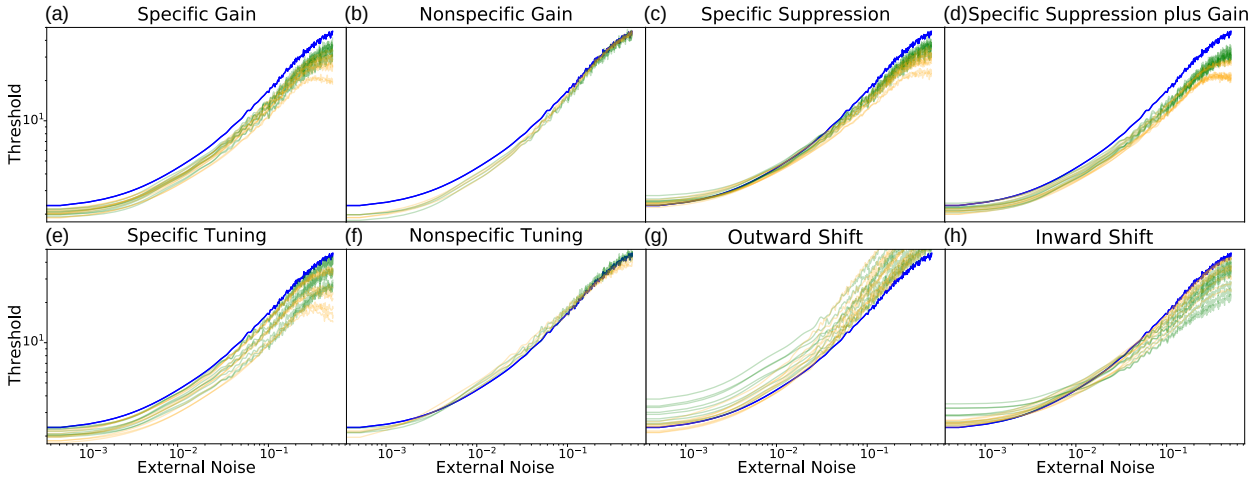

Figure 8: TvN curves are depicted for each mechanism of neural code change. Blue curves correspond to baseline models, orange curves to encoding changes implemented in a dense population (20 channels), and green curves to encoding changes implemented in a sparse population (10 channels).

(a) Specific Gain. The TvN curve produced after applying specific gain is consistently more precise (i.e., lower threshold) than that of the baseline across all levels of external noise. The increased responsiveness of channels around the target leads to steeper slopes in their tuning functions, This improves the channels’ sensitivity to stimulus changes around the target exclusively, without increasing sensitivity to stimulation due to external noise in off-target channels; thus, compared to baseline, the precision is improved across all levels of external noise.

(b) Nonspecific Gain. The TvN curve produced after applying nonspecific gain is initially more precise than that of the baseline at the flat and curved sections, but precision converges with the baseline by the end of the linearly-increasing section. At the flat section, much like specific gain, increased responsiveness leads to steeper slopes along channels surrounding the target, improving decoding precision. At the linearly-increasing section, when the strength of the external noise becomes more comparable to that of the target stimulus, response to noise at the off-target channels is amplified and there isn’t an improved effect compared to baseline.

(c) Specific Suppression. The TvN curve produced after applying specific suppression is higher than baseline (more imprecise) at the flat section, crosses baseline around the curved section, and

it becomes lower than baseline (more precise) at the linearly-increasing section. The general effect of specific suppression is to decrease sensitivity to noise in the off-target channels. The level of suppression determines to what extent this also affects the channels close to the target. With high levels of suppression, the responsiveness of channels surrounding the target is reduced, which reduces their slope and the precision of decoding at low levels of noise. The improvements in precision in the linearly-increasing section are due to the relatively stronger reduction of slopes in off-target channels, which makes them less sensitive to broadband noise. Figure 8c also shows an exception to the general pattern, in which specific suppression increases decoding precision in the flat part of the TvN curve. This happens when suppression is strong enough to reduce the influence of internal neural noise from channels surrounding the target, but weak enough to not disrupt the slope of those channels too strongly.

(d) Specific Suppression plus Gain. The TvN curve produced after applying specific suppression with nonspecific gain is generally more precise than that of the baseline across all levels of external noise. While the basic features are consistent with specific gain and specific tuning, the slope of the linearly-increasing section is more gradual along the TvN curve compared to baseline, while both specific gain and tuning produce a TvN that runs in parallel to the baseline throughout the linearly-increasing section. Mechanistically, the improvement at high external noise levels is due to response reduction in off-target channels, as in specific suppression. To this, non-specific gain adds an improvement at low external noise levels, due to an overall increase in slopes. Whether or not this improvement is enough to bring imprecision to a level lower than baseline depends on the balance between gain and suppression. As the nonspecific gain decreases, the TvN curve becomes more similar to regular specific suppression. This explains some exceptions in Figure 8, in which imprecision is higher than baseline at the lowest levels of external noise.

(e) Specific Tuning. The TvN curve produced after applying specific tuning is consistently more precise than that of the baseline across all levels of external noise. The TvN curves are similar to those obtained with specific gain, and the underlying mechanism is again an increase in slopes for tuning functions surrounding the target. Due to the specificity of this mechanism, the change improves channel sensitivity to stimulus changes around the target exclusively, without increasing sensitivity to stimulation due to external noise in off-target channels. This results in a constant improvement in decoding along the linearly-increasing section of the TvN curve.

(f) Nonspecific Tuning. The TvN curve produced after applying nonspecific tuning is more precise than that of the baseline during the flat section, but less precise during the curved section and the majority of the linearly-increasing section before finally matching the baseline. While the channels near the target have increased their sensitivity by aligning their steepest points to the target (explaining improvement in the flat section), external noise is no longer uniquely beneficial at the target location as it was with specific tuning, and the indiscriminately improved sensitivity to stimuli (including external noise) is detrimental during the linearly-increasing section.

(g) Outward Shift. The TvN curve produced after applying an outward tuning shift is in most cases more imprecise than that of the baseline across all levels of external noise. In the general case, moving channels away from the target decreases the slopes of tuning functions at the target. This in turn decreases the channels' sensitivity to stimuli near the target and reduces precision.

The decreased channel sensitivity applies across all three sections of the curve: adding noise just makes it worse. The single exception among our simulations is one in which an improvement in target decoding is seen at low levels of external noise. We believe this to be an artifact of our simulation parameters. This effect is observed when the shift in tuning is so large that only a channel with the target as its preferred stimulus is left unshifted. The center of the tuning shift and the preferred stimulus of one channel must perfectly match for this to happen. Under such conditions, our assumption of a near-equal variance SDT model fails, because decoding precision is high exactly at the target (a maximum response value corresponds only to the target) but quickly decreases at stimulus values slightly off-target (due to symmetry in the tuning function, a given response value corresponds to two possible stimulus values).

(h) Inward Shift. The TvN curve produced after applying an inward tuning shift is generally more imprecise than that of the baseline during the flat and curved sections, but the precision improves during the linearly-increasing section (much like the pattern produced after applying specific suppression). The drop in decoding precision around the target at low levels of noise happens because inward shift makes the tuning functions of multiple channels similar to that of the channel exactly at the target. More similar responses means that channels to the sides of the target provide more redundant information about the presented stimulus. During the linearly-increasing section, precision increases because the response of channels near the target is increased disproportionately by the high broadband noise compared to distal channels. Inward shift means that less tuning functions cover stimuli away from the target, and thus external noise that matches those stimuli does not produce a response as strong as the response produced for stimuli around the target.

Generally speaking, TvN curves are able to differentiate nonspecific gain, nonspecific tuning, and outward tuning shifts from all other models. Unfortunately, the TvN curves produced by inward tuning shifts are very similar to those produced by specific suppression, and populations affected by specific gain, specific tuning, and specific suppression plus nonspecific gain produce highly similar TvN curves as well.

Threshold vs Stimulus (TvS) Curves

Differences in the number and properties of neurons encoding a particular dimensional value should produce differences in the precision with which that dimensional value can be decoded. In general, decoding from neurons which are more numerous, more finely tuned, and have a larger range of responses (i.e., difference between baseline and maximum firing rate) is more precise. For example, Paradiso (1988) showed that the precision with which orientation can be decoded from a neural population strongly depends on the number of cells encoding such variable. Due to the cortical magnification factor in V1, the number of cells encoding orientation drops with eccentricity, and this drop provides an excellent fit to estimates of orientation thresholds as a function of eccentricity.

The different mechanisms of change in population codes shown in Figure 2 should therefore produce concomitant changes in thresholds measured at different values of the stimulus dimension. An experiment measuring thresholds at many different “pedestal” values of the dimension should produce a Threshold versus Stimulus (TvS) curve that would provide important information about underlying changes in population codes. A TvS curve is also much easier to interpret than many other possible psychophysical functions. Sections of the curve with higher values represent more imprecise decoding estimates, which result from neurons that are relatively fewer in number,

more broadly tuned, or with a smaller response range. Sections of the curve with lower values represent more precise decoding estimates, which result from neurons that are relatively more in number, more finely tuned, or with a larger response range.

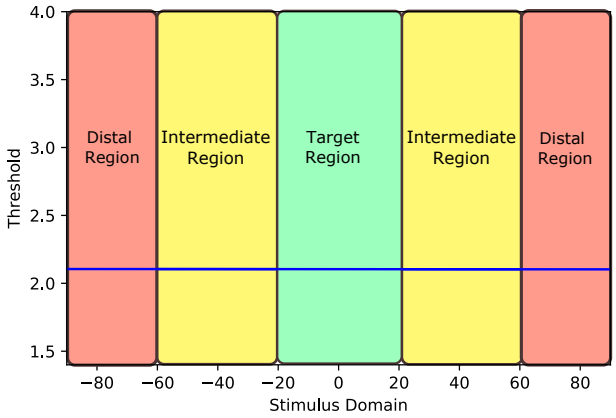

Figure 9: Division of the stimulus dimension into regions as a function of their distance from the target stimulus. The division is rather arbitrary, and we propose it here mostly as a way to interpret the simulated TvS curves shown in Figure 10.

As shown in Figure 9, the sections of the TvS function are loosely categorized based on their distance from the target. Given that our simulations focus on circular dimensions, such as orientation, stimuli farthest from the target are shown on the left and right side: the distal section. The target itself and nearby stimuli fall under the target section. The remaining cases fall under the intermediate section. The TvS functions were all acquired in the absence of external noise, which means the target section should correspond to the initial flat region of the TvN curves.

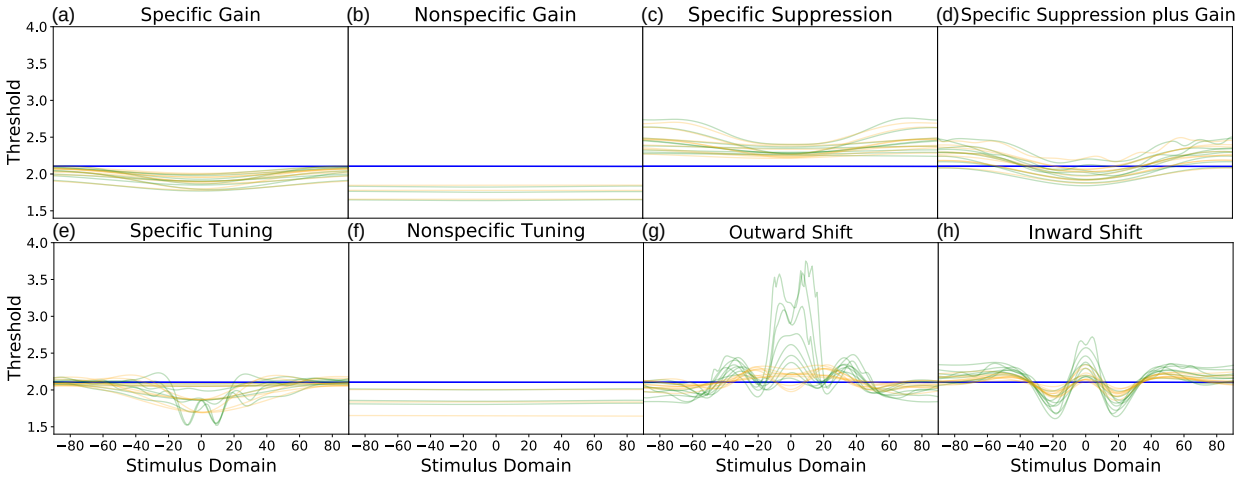

Figure 10: TvS curves are depicted for each mechanism of neural code change. Blue curves correspond to baseline models, orange curves to encoding changes implemented in a dense population (20 channels), and green curves to encoding changes implemented in a sparse population (10 channels).

(a) Specific Gain. The TvS function produced after applying specific gain is generally more precise than that of the baseline, but gradually approaches the baseline precision as the effect dissipates toward the distal section. The sensitivity of the channels are increased most in the target region (due to increased slopes of tuning functions in that region) and least in the distal region, explaining the effect.

(b) Nonspecific Gain. The TvS function produced by applying nonspecific gain is always more precise than that of the baseline. The channel slopes are increased indiscriminately, allowing the effect observed in specific gain to be applied across the stimulus domain. The level of gain directly determines the degree of the increase in decoding precision. Without the interference of external noise, the increase in responsiveness is useful for all values of the stimulus domain.

(c) Specific Suppression. The TvS function produced by applying specific suppression is always more imprecise than that of the baseline. The imprecision increases toward the distal sections. Specific suppression produces flatter slopes across tuning functions of all channels, reducing decoding precision, but the effect is stronger the farther away a channel is from the target.

(d) Specific Suppression plus Gain. The TvS function produced by applying specific suppression plus gain is more precise than the baseline at the target and intermediate sections, but more imprecise in the distal section. The shape of the TvS function matches specific suppression exactly, but its vertical position is determined by the level of nonspecific gain: adding enough nonspecific gain can lead to a TvS curve exactly like that of specific gain, and adding too little leads to a TvS curve exactly like that of specific suppression. However, if enough gain was modeled to match or surpass baseline along all sections, the mechanism would not truly qualify as involving suppression.

(e) Specific Tuning. The TvS function produced after applying specific tuning is more precise than baseline in the target section. Within the target section especially, the slopes of tuning functions are increased, which accounts for the corresponding improvements in precision. In 50% of the simulations, specific tuning caused the TvS function to increase imprecision above the baseline during the intermediate and/or distal sections. Finding such “shoulders” in the TvS function would differentiate specific tuning from specific gain, but not finding them is insufficient to differentiate this mechanism from specific gain (this comparison is particularly important, as TvN curves for these two models are also indistinguishable). For the rest of the simulations, five of them (27.8%) involved a change in tuning so small that no change was visible in the TvS curve except for a very small increase in performance for all values of the stimulus. The last three cases involved curves similar to those observed for specific gain. Thus, there are cases in which the TvS function for specific tuning mimics that observed for specific gain.

(f) Nonspecific Tuning. The TvS function produced after applying nonspecific tuning is more precise than baseline across all sections. As another flat line, the function is seemingly identical to the nonspecific gain function. The uniform precision increments are due to the indiscriminate increase in slopes across tuning functions. As shown in the corresponding TvN curve, the improvement only applies without the interference of external noise.

(g) Outward Shift. The TvS function produced after applying an outward tuning shift is generally less precise than baseline at the target section, and the precision generally improves at the distal section. Shifting of tuning functions away from the target tends to concentrate slopes in the farthest area, where more precise decoding is made possible. The intermediate channels, which move the most, also cannot assist with decoding stimuli in the target region. The redistribution of slopes seems beneficial to decoding only starting at intermediate distances from the target (i.e., around +/-50), but is detrimental to decoding at values closer to the target.

(h) Inward Shift. The TvS function produced after applying an inward tuning shift is less precise at the target region, more precise in the intermediate region, and again less precise in regions farthest from the target. Shifting of tuning functions towards the target tends to concentrate slopes in the intermediate area, with the consequence of a lower concentration of slopes in the region closest to and farthest from the target.

Generally speaking, TvS functions are able to differentiate specific suppression, specific suppression plus gain (when parameters do not make either of the two mechanisms dominate), outward shift, and inward shift from all other models. Unfortunately, the TvS functions for specific gain and tuning were too similar to each other to distinguish them, and the functions for nonspecific gain and nonspecific tuning were practically identical. However, TvS functions allow researchers to still narrow down the type of change in encoding to either a specific or nonspecific mechanism. If a nonspecific mechanism is found, then gain and tuning can be further separated by using TvN functions, as shown in the previous section.

Discussion

We performed a large number of simulations to determine what types of neural encoding changes could be differentiated through psychophysical threshold experiments and IEM-estimated population responses. A summary with the most common patterns of results observed in our simulations is shown in Table 2, where unique predictions that are diagnostic of a specific model are highlighted in color. The results suggest that no single experiment allows one to distinguish between all mechanisms of encoding change. A psychophysical TvS curve is the most informative piece of data, distinguishing four of the eight types from all other mechanisms. This was followed by the TvN curve, which distinguished three types. Population response estimates from IEM, as they are most often obtained (i.e., through shifted average), could not distinguish any of the mechanisms, although they could distinguish between the two general classes of gain and tuning. However, a combination of a psychophysical and a neuroimaging study would allow researchers to distinguish all eight of the mechanisms under study here. While a TvS curve distinguishes specific suppression, specific suppression plus gain (as long as both mechanisms have a balanced contribution), outward tuning shift, and inward tuning shift, estimation and analysis of population responses estimated through IEM can distinguish between all the remaining mechanisms of encoding change (specific and nonspecific gain and tuning), in addition providing information of exactly where in the brain encoding changes are observed. Thus, our results suggest that the goal of making valid inferences about mechanisms of encoding change from indirect measures could be achieved through the combination of both neuroimaging and psychophysical data.

One interpretive issue that researchers might encounter is that IEM reveals changes in population responses linked to a variable of interest (e.g., attention or learning) in multiple brain regions of interest (ROIs). Because such changes are unlikely to be the same across ROIs, the

question remains as to what brain areas are linked to the behavioral changes observed through psychophysics. How can behavioral and neural data be better integrated in such a case? One solution would be to perform model selection using the psychophysical data, either through comparison of qualitative data patterns (i.e., Table 2) or quantitative model fit and selection. Then, the selected behavioral model can be used to generate predictions of what the results of IEM should look like in a region containing the key neural populations providing information for behavior. Such predictions would be similar to the simulated results shown in the bottom panel of Figure 3. They can be compared against population responses estimated through IEM, to determine which ROI provides estimates matching the behavioral model best. The best match would be a good candidate for the region providing most information for behavior. The behavioral model can also be used to directly generate predictions about population responses that are included in a regression analysis with neuroimaging data as the dependent variable. This is equivalent to estimating the weights in a linear measurement model (or mixing model, see Equation 8) describing how much each neural channel contributes to the activity in a given measurement (e.g., voxel or EEG channel). The reader might recognize this as traditional forward encoding modeling (see Soto and Ashby, 2023), which would result in a set of voxels that are well-explained by the behavioral model and that may not be restricted to a particular ROI.

For a researcher focused on collecting only behavioral data, all mechanisms of encoding change cannot be distinguished, but the combination of two experiments (TvS and TvN) would considerably narrow down the range of possibilities. In addition to provide unique predictions for four mechanisms, finding one of the other two nondiagnostic (TvS patterns shown in Table 2 (in black) would allow one to determine whether the underlying mechanism is specific to the target stimulus (--= pattern) or nonspecific (--- pattern). If the pattern is nonspecific, addition of a TvN curve would distinguish between nonspecific gain and nonspecific tuning. If the pattern is specific, then an increase in thresholds for intermediate and/or distal regions of the TvS curve is indicative of a specific tuning mechanism. In a minority of cases (3/18 of the simulations), a specific tuning mechanism produced a TvS curve mimicking that of a specific gain mechanism, and thus both cannot be separated unless other sources of information are taken into account.

Model Name	PR	TvN	TvS
Specific Gain	+/-/+	---	--=
Nonspecific Gain	+/+/+	**--=**	---
Specific Suppression Plus Gain	+/-/+	---	**-=+**
Specific Suppression	-/-/+	+=-	**+++**
Specific Tuning	-/-/+	---	--=
Nonspecific Tuning	-/-/+	**-+=**	---
Outward Shift	-/-/+	**+++**	**++-**
Inward Shift	+/+/+	+=-	**+-+**

Table 2: Most common patterns of results observed in our simulations, with symbols representing changes from baseline. Values higher, equal, and lower than baseline are represented by +, =, and -, respectively. Symbols in the Population Response (PR) column represent the average difference with baseline in the three parameters of a curve fitted to the population response to the target: curve height, curve width, and peak curvature, in that order. Symbols in the Threshold versus Noise (TvN) column represent difference with baseline in the flat, curved, and linearly increasing sections of the curve, in that order. Symbols in the Threshold versus Stimulus (TvS) column represent difference with baseline in the area around the target, in the intermediate section, and the distal section, in that order.

While TvN and TvS alone could not dissociate every mechanism (i.e., specific gain and specific tuning produced qualitatively identical patterns), additional psychophysical studies may be able to differentiate between the two mechanisms. For example, assuming that the overall level of activity elicited by a stimulus determines its relative salience, producing bottom-up attention, then a target under specific gain should capture attention more easily in visual search tasks than a target under specific tuning (Soto et al., 2021). More generally, at the heart of our approach is the idea that no single study allows one to infer the correct mechanism of encoding change underlying some behavior of interest. Rather, a combination of multiple studies, all linked together through the same model, provides a much stronger approach to the problem.

Re-interpreting results in the literature

As indicated in the introduction, we have focused here on those mechanisms of encoding change that have been proposed to improve task performance in the previous literature, making them candidates to explain the effects of learning and attention on perceptual processing. Although no prior experiment has attempted to distinguish between multiple mechanisms using both TvS and TvN functions, several experiments have gathered thresholds either at the target stimulus (i.e., the stimulus involved in learning or attention), across different values of the stimulus dimension (i.e., similar to a TvS function), or across different levels of external noise (i.e., a TvN function). The results of some of these studies can be re-interpreted in the light of our current simulations.

For example, several studies have shown that aversive Pavlovian conditioning involving a particular stimulus (the conditioned stimulus, or CS+, which is paired with an aversive stimulus, such as electric shock) produce increments in thresholds for that stimulus (Shalev et al., 2018; Laufer and Paz, 2012; Resnik et al., 2011). Figure 10 shows that only a few mechanisms can produce this increase in thresholds at the target: specific suppression, outward shift, or inward shift. However, other evidence suggest that, among these candidates, the most likely mechanism is one of inward shift. First, neurophysiological studies in auditory aversive conditioning with rodents have shown evidence that individual neurons shift their preferred stimulus towards the CS+ after training (for reviews, see Weinberger 2007, 2011). Second, there are multiple reports that an aversive CS+ captures attention in search tasks (Van Damme et al., 2004; Koster et al., 2005, 2004; Notebaert et al., 2011). A reasonable assumption is that bottom-up attentional capture depends on the overall level of neural activity that the CS+ produces in comparison with concurrently-presented stimuli that compete for attention (Soto et al., 2021). If we think of that overall level of neural activity as the result of both the number of neurons selective for the CS+ as well as their firing rates, we see that both suppression and outward shift are mechanisms likely to reduce the overall level of neural activity produced by the CS+, whereas an inward shift is likely to produce the opposite effect. In sum, both neurophysiological and psychophysical data suggest that the most likely change in stimulus encoding produced by aversive conditioning is an inward shift towards the CS+. This hypothesis could be easily tested by estimating TvS functions from participants before and after conditioning. Based on our results, we would predict that the post-learning TvS function would look like the inward shift function in Figure 10.

Another form of learning known to produce changes in dimension discriminability is category learning, although such changes have usually been quantified using a measure of sensitivity (i.e., d') rather than sensitivity thresholds. In particular, several studies have shown that categorization training produces increased discriminability along the category-relevant stimulus dimension (Folstein et al., 2012, 2013, 2014; Goldstone, 1994; Notman et al., 2005; Op de Beeck et al., 2003; Van Gulick and Gauthier, 2014), an effect that can be sometimes stronger for stimuli that cross the category boundary (Goldstone, 1994; Notman et al., 2005). Such results are compatible with find-

ings from a recent IEM study (O'Bryan et al., 2024), where category learning was accompanied by changes in encoding around the category boundary that could be explained by a specific gain mechanism. The psychophysical results are less compatible with another IEM study (Ester et al., 2020) supporting the operation of an outward tuning shift mechanism during categorization of oriented gratings. In particular, Figure 10 shows that the outward shift mechanism found by Ester et al. in most cases would lead to a reduction in stimulus discriminability (i.e., increased thresholds, or decoding imprecision) around the bound (which in this case corresponds to the target at zero). Improvements in discriminability should be reliably observed only for stimuli that are relatively far from the bound. In line with this idea, Van Gulick and Gauthier (2014) report some evidence of a stronger effect on discriminability away from the category bound than at the category bound. On the other hand, Goldstone (1994) reports the opposite pattern of results, and others (Folstein et al., 2012, 2013) report equivalent improvements in discriminability across the dimension.

As pointed out by O'Bryan et al. (2024), one way to harmonize these disparate results is by hypothesizing that the area of a stimulus dimension showing the strongest encoding modulation changes with training, in line with other results in the category learning literature (e.g., Soto et al., 2013, 2016). Early in training, modulation would be focused around the bound, but this focus would shift away from the bound as training advances and participants shift toward a more automatic categorization strategy. Still, what all psychophysical studies have in common is that they report an increment in discriminability for stimuli around the category bound, which is difficult to explain by an outward shift mechanism. It is possible that a precisely-tuned outward shift can produce this result, if the shifts place the slopes of several tuning functions around zero. In any case, an outward shift mechanism predicts a loss in discriminability in some areas of the dimension, which has not yet been observed. Because a tuning shift necessarily increases discriminability in some areas at the cost of a reduction in discriminability in other areas, we predict that if an outward shift is the mechanism underlying perceptual effects of categorization, then a fine-grained TvS curve should reveal areas in which the category-relevant dimension shows a reduction in stimulus discriminability.

An additional complication is that the precise mechanism of encoding change produced by categorization learning might depend on properties of the neural population encoding a particular stimulus dimension. More precisely, both O'Bryan et al. (2024) and Ester et al. (2020) used orientation of lines and gratings as their stimuli, known to be encoded in early visual cortex through tuning functions similar to those used in the present study. On the other hand, the majority of the psychophysical research has used either highly complex shape and object stimuli (Folstein et al., 2012, 2013, 2014; Op de Beeck et al., 2003; Van Gulick and Gauthier, 2014) or simple dimensions other than orientation (Goldstone, 1994; Notman et al., 2005). The tuning functions used by the brain to encode such dimensions might be different than what is represented by the standard population model used here. For example, face features could be encoded through sigmoidal tuning functions ((McKone et al., 2014; Soto et al., 2020, 2021)). Using computational modeling and visual adaptation, it has been found that the effects of categorization on perception of face identities along the category-relevant dimension (Goldstone and Steyvers, 2001; Soto and Ashby, 2019, 2015) can be best explained using a specific gain mechanism (Soto et al., 2020). It is currently unknown exactly how the complex stimuli used in some studies are encoded, but encoding that is different from that of orientation might be at the heart of the results obtained with such dimensions. A better understanding of the results from recent IEM studies (Ester et al., 2020; O'Bryan et al., 2024) could be achieved by obtaining TvS and TvN curves across multiple stages of category learning, in a categorization task using oriented grating stimuli.

Other studies have reported TvN curves to characterize the effects of learning in terms of psychophysical observer models (Lu and Dosher, 2008). The results of such studies can be

re-interpreted in terms of the encoding/decoding observer model studied here (Figure 1). For example, perceptual learning results in TvN curves that drop from baseline at all levels of external noise (Dosher and Lu, 1998; Gold et al., 1999; Xie and Yu, 2018). This result is consistent with multiple mechanisms of encoding change (see Figure 8): specific gain, specific tuning, and specific suppression plus gain. Interestingly, all these mechanisms are in line with the well-known stimulus specificity of perceptual learning. To distinguish among these different potential mechanisms, more information can be obtained through a TvS curve, but additional steps might be required to differentiate between specific gain and tuning (see section *Recommendations for researchers* below).

Finally, There are multiple studies that have estimated TvN curves under different attentional demands. The study that is closest to our work was performed by Ling et al. (2009a), who specifically sought out to dissociate between nonspecific gain and specific suppression plus gain as mechanisms of attention. The authors chose those two mechanisms because they would specifically reduce thresholds in the early (internal noise suppression) and late (external noise suppression) parts of the TvN curve, respectively. Our simulations have confirmed that, among all the mechanisms of encoding change studied, nonspecific gain is the only one that seems to uniquely affect internal noise. On the other hand, there is no mechanism that can uniquely reduce external noise. Specific suppression plus gain can uniquely affect external noise if the two mechanisms are finely tuned, as the reduction in thresholds in the early part of the curve produced by nonspecific gain (see Figure 8b) can be chosen so that it perfectly counteracts the increment in those thresholds produced by suppression (see Figure 8c), leaving only a reduction in thresholds at the end part of the curve. In most of our simulations, we find that the TvN curve drops below baseline at low levels of external noise (see Figure 8d).

The results reported by Ling et al. suggest that the mechanism of encoding change underlying spatial attention is nonspecific gain, which is in line with other reports (Lu and Dosher, 1998b). On the other hand, feature-based attention produced a drop in the TvN function at all levels of external noise, which is consistent with multiple mechanisms not considered by the authors.

Recommendations for researchers

When measuring psychophysical thresholds, we recommend using sensitivity thresholds (a stimulus value related to a specific value of d') rather than the more commonly-used accuracy thresholds (a stimulus value related to a specific proportion of correct responses). While accuracy thresholds would be contaminated with bias, sensitivity thresholds take response bias into account. There are currently methods available to estimate sensitivity thresholds using both yes/no and 2AFC tasks (Lesmes et al., 2015). An important theoretical reason to prefer sensitivity thresholds is that the theory links them directly to decoding precision in the encoder/decoder observer model.

Also, we found that obtaining estimates at the target stimulus, with or without the inclusion of external noise (i.e., TvN curve), is insufficient to distinguish between all of the mechanisms of encoding change. Thus, we recommend gathering estimates for stimuli surrounding the target as well (i.e., TvS curve), to elucidate the neural mechanisms at work.

Using adaptive psychophysical procedures, such as QUEST+ (Watson, 2017), it is possible to obtain a reliable threshold estimate in about 100-150 trials, taking around five to ten minutes. It is difficult to provide fixed rules on how many thresholds should be obtained to estimate changes in a TvS curve, but a researcher using IEM might want to obtain a threshold for each stimulus presented during the IEM experiment. In an unpublished study, we found that a TvS with nine thresholds was enough to distinguish among most mechanisms of encoding change produced by category learning. Traditional studies have obtained about eight thresholds to characterize the

TvN curve (Dosher and Lu, 1998, 2000; Lu and Dosher, 2000, 1998a), but more efficient methods exist to estimate the curve (Lesmes et al., 2006). A caveat is that each curve must be estimated both at a baseline and during operation of the mechanism under study (e.g., under attention or after learning). While psychophysical experiments can be relatively long, they are low-cost and their inclusion together with IEM studies would offer a large payoff: being able to reach more definitive conclusions about the precise mechanisms of encoding change at work in a particular situation of interest. In our experience, even stronger conclusions can be reached if model fit and selection are used for data analysis (Cavagnaro et al., 2013; Myung et al., 2007; Zucchini, 2000).

No singular experiment could distinguish every possible change in neural codes. We recommend acquiring thresholds as a function of multiple factors throughout multiple experiments, but if manipulating only one factor is possible, then measuring thresholds along the stimulus domain seems to be very effective. On the other hand, if a researcher is interested in telling apart a specific mechanism of encoding change from all others, then it could be wiser to obtain a TvN curve instead. For example, either nonspecific gain or nonspecific tuning can be separated from all other mechanisms using a TvN curve, but not a TvS curve. When faced with uncertainty between two candidate models, a good approach might be to fit the models to the observed data and perform formal model selection (Cavagnaro et al., 2013; Myung et al., 2007; Zucchini, 2000), although this would be a computationally intensive procedure employing the Monte Carlo technique used here to obtain thresholds from the model.

It is also important to note that these models were differentiated under specific assumptions about encoding and decoding; therefore, we recommend that researchers carefully consider what assumptions they will be able to justify (i.e., the use of an “aware” optimal decoder may not be justifiable in adaptation experiments; see Series et al., 2009).

Here, we have focused mostly on basic mechanisms of encoding change that alter a single aspect of the tuning curve. Caution should be maintained when combining basic mechanisms into more complex ones. In our simulations, specific suppression plus gain was consistently an interpolation between specific suppression and specific gain. If a researcher is particularly interested in testing combinations of basic mechanisms of encoding change, the parameters need to be balanced to avoid ambiguity. In our example, specific suppression plus a very small gain looks very similar to ordinary specific suppression, whereas large gain with small specific suppression looks very similar to ordinary nonspecific gain.

Limitations

A number of assumptions were made in this simulation work, which are listed in what we perceive is their order of importance in Table 3. First, because there is inherent ambiguity in linking the neural and psychophysical levels, such that any changes in thresholds could be attributable to either encoding or decoding changes (Gold and Ding, 2013), we started by assuming an optimal decoder. This is a rather strong assumption, but is also standard in the prior literature (e.g., Dakin et al., 2005; Deneve et al., 1999; Ling et al., 2009a; May and Solomon, 2015; Series et al., 2009; Paradiso, 1988), where it has proven to be useful. In addition, biologically plausible mechanisms for optimal decoding in the brain have been proposed in the literature (Deneve et al., 1999). Second, we also assumed independent Poisson neural noise, which facilitates maximum likelihood estimation and is also a common assumption in the literature (e.g., Dakin et al., 2005; Deneve et al., 1999; Ling et al., 2009a; May and Solomon, 2015; Series et al., 2009; Paradiso, 1988). Third, we assumed that decoding noise is similar in neighboring areas of the stimulus dimension. Once again, this is a common assumption in the literature linking neural encoding with psychophysics (Dakin et al., 2005; Ling et al., 2009a), made mostly for convenience as it substantially reduces

the computational cost of simulations. We believe that this assumption also seems valid, as there is little reason to expect large differences in decoding precision in a small area of the stimulus domain. Fourth, we assumed a bell-shaped tuning function, which as far as we know is the only type of function used in research applying IEM (e.g., Garcia et al., 2013; Ester et al., 2020; Serences et al., 2009; Sprague et al., 2018, 2019; Gardner and Liu, 2019; Liu et al., 2018; Brouwer and Heeger, 2009). For other stimulus dimensions, such as those characterized by monotonically increasing or decreasing tuning functions, new simulations would be required. Finally, we assumed a homogeneous encoding population for the baseline condition. This is a common assumption in IEM, and we don't think that changes in this assumption would change any of our conclusions. This is because, regardless of what baseline is assumed, changes in encoding relative to that baseline should result in similar changes in decoding precision relative to the baseline.

As we have indicated, some of the assumptions in Table 3 are rather strong. However, we do not believe that they are stronger than the assumptions that IEM requires about encoding (e.g., homogeneous population codes, normal neural noise) and about the link between neural activity and neuroimaging measures (a linear measurement model with additive normal noise at each measurement, and independent across measurements; see Soto and Ashby 2023; Soto et al. 2018; Van Bergen et al. 2015). We believe that the weaknesses of each approach can be overcome by combining different sources of data (which require different sets of assumptions) within a single modeling framework. In the future, a powerful methodology would involve using both psychophysics and neuroimaging data together to infer changes in encoding, perhaps using hierarchical Bayesian modeling, in which multiple types/sources of data for each participant can be used simultaneously to make inferences about a single set of model parameters (Palestro et al., 2018).

Assumptions
Optimal MLE Decoding
Independent Poisson Neural Noise
Equivalent Noise Variance in Neighboring Areas of Stimulus Dimension
Bell-Shaped Tuning Function
Homogeneous Encoding Baseline*

Table 3: List of assumptions made in the present simulation work, ordered from strongest to weakest.

We chose an optimal decoder that would be considered "aware" (Series et al., 2009). That is, we assumed that the decoder had complete knowledge of the statistics of each encoding model, before and after application of a given mechanism of encoding change. It is known that this assumption is unlikely to hold for neural code changes underlying adaptation (Series et al., 2009), perhaps because it results from transient environmental events, and the decoder may not have opportunity to learn and adapt to those changes. However, we think that an aware decoder is a good assumption for situations involving encoding changes that are predictable due to extensive prior experience, such as those produced by learning and other cognitive mechanisms. For example, aware decoding has been widely used in previous modeling of attentional mechanisms (e.g., Itti et al., 1998; Lee et al., 1999; Ling et al., 2009a; Nakahara et al., 2001; Pestilli et al., 2009), with the underlying assumption being that attentional control processes have precise information about the results of attention in population codes, and that this information is used during decoding.

In addition, although optimal decoding through MLE has been widely used, other decoding schemes are possible. For example, it is possible to use simpler linear decoders or, if the popu-

lationdensely covers the stimulus space, a max response decoder. However, MLE is biologically plausible (Deneve et al., 1999) and provides a more meaningful benchmark due to it being optimal for the task.

Conclusion

Psychophysical thresholds can help researchers infer the mechanisms of encoding change due to cognitive states of interest, and thus are a great source of constraints for researchers using IEM to make inferences about encoding rather than population responses. Because any quantifiable stimulus dimension can be used to produce a TvS or TvN function, this approach can be applied to a variety of high- and low-level stimuli to answer a plethora of questions in vision neuroscience.

References

Antov, M. I., Plog, E., Bierwirth, P., Keil, A., and Stockhorst, U. (2020). Visuocortical tuning to a threat-related feature persists after extinction and consolidation of conditioned fear. *Scientific Reports*, 10:3926.

Baldassi, S. and Verghese, P. (2005). Attention to locations and features: Different top-down modulation of detector weights. *Journal of Vision*, 5(6):7–7.

Brouwer, G. J. and Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, 29(44):13992–14003.

Byers, A. and Serences, J. T. (2014). Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. *Journal of Neurophysiology*, 112(5):1217–1227.

Cavagnaro, D. R., Myung, J. I., and Pitt, M. A. (2013). Mathematical modeling. In *The Oxford Handbook of Quantitative Methods*, volume 1, pages 438–453. Oxford University Press, New York, NY.

Dakin, S. C., Mareschal, I., and Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research*, 45(24):3027–3049.

Deneve, S., Latham, P. E., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745.

Dosher, B. A. and Lu, Z. L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, 95(23):13988–13993.

Dosher, B. A. and Lu, Z. L. (2000). Noise exclusion in spatial attention. *Psychological Science*, 11(2):139–146.

Ester, E. F., Anderson, D. E., Serences, J. T., and Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, 25(5):754–761.

Ester, E. F., Sprague, T. C., and Serences, J. T. (2020). Categorical biases in human occipitoparietal cortex. *Journal of Neuroscience*, 40(4):917–931.

Folstein, J. R., Gauthier, I., and Palmeri, T. J. (2012). Not all morph spaces stretch alike: How category learning affects object discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4):807–802.

Folstein, J. R., Palmeri, T. J., and Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4):814–823.

Folstein, J. R., Palmeri, T. J., and Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: the case of shape and motion. *Cognition*, 5:1394.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex*, 16(11):1631–1644.

Garcia, J. O., Srinivasan, R., and Serences, J. T. (2013). Near-real-time feature-selective modulations in human cortex. *Current Biology*, 23(6):515–522.

Gardner, J. L. and Liu, T. (2019). Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *eNeuro*, 6(2):e0363–18.2019.

Gold, J., Bennett, P. J., and Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature*, 402(6758):176–178.

Gold, J. I. and Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, 103:98–114.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178–200.

Goldstone, R. L. and Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1):116.

Harrison, W. J., Bays, P. M., and Rideaux, R. (2023). Neural tuning instantiates prior expectations in the human visual system. *Nature Communications*, 14(1):5320.

Itti, L., Braun, J., Lee, D. K., and Koch, C. (1998). Attentional modulation of human pattern discrimination psychophysics reproduced by a quantitative model. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems*, volume 11, pages 789–795. MIT Press.

Koster, E., Crombez, G., Van Damme, S., Verschuere, B., and De Houwer, J. (2005). Signals for threat modulate attentional capture and holding: Fear-conditioning and extinction during the exogenous cueing task. *Cognition & Emotion*, 19(5):771–780.

Koster, E. H. W., Crombez, G., Van Damme, S., Verschuere, B., and De Houwer, J. (2004). Does Imminent Threat Capture and Hold Attention? *Emotion*, 4(3):312–317.

Laufer, O. and Paz, R. (2012). Monetary Loss Alters Perceptual Thresholds and Compromises Future Decisions via Amygdala and Prefrontal Networks. *Journal of Neuroscience*, 32(18):6304–6311.

Lawson, R. P., Clifford, C. W. G., and Calder, A. J. (2011). A real head turner: Horizontal and vertical head directions are multichannel coded. *Journal of Vision*, 11(9):17–17.

Lee, D. K., Itti, L., Koch, C., and Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2(4):375–381.

Lehky, S. R., Sereno, M. E., and Sereno, A. B. (2013). Population coding and the labeling problem: extrinsic versus intrinsic representations. *Neural Computation*, 25(9):2235–2264.

Lesmes, L. A., Jeon, S. T., Lu, Z. L., and Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick tvc method. *Vision research*, 46(19):3160–3176.

Lesmes, L. A., Lu, Z.-L., Baek, J., Tran, N., Dosher, B. A., and Albright, T. D. (2015). Developing Bayesian adaptive methods for estimating sensitivity thresholds (d') in Yes-No and forced-choice tasks. *Frontiers in Psychology*, 6.

Ling, S., Jehee, J. F. M., and Pestilli, F. (2015). A review of the mechanisms by which attentional feedback shapes visual selectivity. *Brain Structure and Function*, 220(3):1237–1250.

Ling, S., Liu, T., and Carrasco, M. (2009a). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, 49(10):1194–1204.

Ling, S., Pearson, J., and Blake, R. (2009b). Dissociation of Neural Mechanisms Underlying Orientation Processing in Humans. *Current Biology*, 19(17):1458–1462.

Liu, T., Cable, D., and Gardner, J. L. (2018). Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width. *The Journal of Neuroscience*, 38(2):398–408.

Lu, Z.-L. and Dosher, B. (1998a). External Noise Distinguishes Attention Mechanisms. *Elsevier*, 38(9):1183–1198.

Lu, Z.-L. and Dosher, B. (2013). *Visual Psychophysics: From Laboratory to Theory*. The MIT Press.

Lu, Z. L. and Dosher, B. A. (1998b). External noise distinguishes attention mechanisms. *Vision Research*, 38(9):1183–1198.

Lu, Z. L. and Dosher, B. A. (2000). Spatial attention: Different mechanisms for central and peripheral temporal precues? *Journal of Experimental Psychology: Human Perception and Performance*, 26(5):1534–1548.

Lu, Z. L. and Dosher, B. A. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychological Review*, 115(1):44–82.

Ma, W. J. (2010a). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, 50(22):2308–2319.

Ma, W. J. (2010b). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, 50(22):2308–2319.

Martinez-Trujillo, J. C. and Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14(9):744–751.

Martinez-Trujillo, J. C. and Treue, S. (2005). The feature similarity gain model of attention: Unifying multiplicative effects of spatial and feature-based attention. In Itti, L., Rees, G., and Tsotsos, J. K., editors, *Neurobiology of attention*, pages 300–304. Elsevier.

May, K. A. and Solomon, J. A. (2015). Connecting psychophysical performance to neuronal response properties I: Discrimination of suprathreshold stimuli. *Journal of Vision*, 15(6):8–8.

McKone, E., Jeffery, L., Boeing, A., Clifford, C. W., and Rhodes, G. (2014). Face identity aftereffects increase monotonically with adaptor extremity over, but not beyond, the range of natural faces. *Vision Research*, 98:1–13.

Myung, I. J., Pitt, M. A., and Kim, K. (2007). Model evaluation, testing and selection. In Lambert, K. and Goldstone, R., editors, *Handbook of Cognition*, pages 422–436. Sage.

Nakahara, H., Wu, S., and Amari, S. (2001). Attention modulation of neural tuning through peak and base rate. *Neural Computation*, 13(9):2031–2047.

Notebaert, L., Crombez, G., Van Damme, S., De Houwer, J., and Theeuwes, J. (2011). Signals of threat do not capture, but prioritize attention: a conditioning approach. *Emotion*, 11:81–89.

Notman, L. A., Sowden, P. T., and Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, 95(2):B1–B14.

O'Bryan, S. R., Jung, S., Mohan, A. J., and Scolari, M. (2024). Category learning selectively enhances representations of boundary-adjacent exemplars in early visual cortex. *Journal of Neuroscience*, 44(3):e1039232023.

Op de Beeck, H. P., Wagemans, J., and Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, 132(4):491–511.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z. L., Steyvers, M., and Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, 84:20–48.

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics*, 58(1):35–49.

Pelli, D. G. (1981). *Effects of visual noise*. Ph.D. Dissertation, Cambridge University.

Pestilli, F., Ling, S., and Carrasco, M. (2009). A population-coding model of attention's influence on contrast response: Estimating neural effects from psychophysical data. *Vision Research*, 49(10):1144–1153.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132.

Pouget, A., Dayan, P., and Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, 26(1):381–410.

Pouget, A., Zhang, K., Deneve, S., and Latham, P. E. (1998). Statistically Efficient Estimation Using Population Coding. *Neural Computation*, 10(2):373–401.

Resnik, J., Sobel, N., and Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. *Nature Neuroscience*, 14(6):791–796.

Salinas, E. and Abbott, L. F. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, 1(1):89–107.

Serences, J., Saproo, S., Scolari, M., Ho, T., and Muftuler, L. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, 44(1):223–231.

Series, P., Stocker, A. A., and Simoncelli, E. P. (2009). Is the homunculus "aware" of sensory adaptation? *Neural Computation*, 21(12):3271–3304.

Seung, H. S. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753.

Shalev, L., Paz, R., and Avidan, G. (2018). Visual Aversive Learning Compromises Sensory Discrimination. *The Journal of Neuroscience*, 38(11):2766–2779.

Song, I. and Keil, A. (2014). Differential classical conditioning selectively heightens response gain of neural population activity in human visual cortex. *Psychophysiology*, 51(11):1185–1194.

Soto, F. A. and Ashby, F. G. (2015). Categorization training increases the perceptual separability of novel dimensions. *Cognition*, 139:105–129.

Soto, F. A. and Ashby, F. G. (2019). Novel representations that support rule-based categorization are acquired on-the-fly during category learning. *Psychological Research*, 83(3):544–566.

Soto, F. A. and Ashby, F. G. (2023). Encoding models in neuroimaging. In Dzhafarov, E. N., Ashby, F. G., and Colonius, H., editors, *New Handbook of Mathematical Psychology: Volume 3: Perceptual and Cognitive Processes*, volume 3 of *Cambridge Handbooks in Psychology*, pages 421–472. Cambridge University Press, Cambridge.

Soto, F. A., Bassett, D. S., and Ashby, F. G. (2016). Dissociable changes in functional network topology underlie early category learning and development of automaticity. *NeuroImage*, 141:220–241.

Soto, F. A., Escobar, K., and Salan, J. (2020). Adaptation aftereffects reveal how categorization training changes the encoding of face identity. *Journal of Vision*, 20(10):18.

Soto, F. A. and Narasiwodeyar, S. (2023). Improving the validity of neuroimaging decoding tests of invariant and configural neural representation. *PLOS Computational Biology*, 19(1):e1010819.

Soto, F. A., Stewart, R. A., Hosseini, S., Hays, J. S., and Beevers, C. G. (2021). A computational account of the mechanisms underlying face perception biases in depression. *Journal of Abnormal Psychology*, 130(5):443–454.

Soto, F. A., Vucovich, L. E., and Ashby, F. G. (2018). Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. *PLOS Computational Biology*, 14(10):e1006470.

Soto, F. A., Waldschmidt, J. G., Helie, S., and Ashby, F. G. (2013). Brain activity across the development of automatic categorization: A comparison of categorization tasks using multivoxel pattern analysis. *NeuroImage*, 71:284–297.

Sprague, T. C., Adam, K. C. S., Foster, J. J., Rahmati, M., Sutterer, D. W., and Vo, V. A. (2018). Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *eNeuro*, 5(3):ENEURO.0098–18.2018.

Sprague, T. C., Boynton, G. M., and Serences, J. T. (2019). The importance of considering model choices when interpreting results in computational neuroimaging. *eNeuro*, 6(6):e0196–19.2019.

Sprague, T. C. and Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, 16(12):1879–1887. Number: 12 Publisher: Nature Publishing Group.

Stegmann, Y., Keil, A., and Wieser, M. J. (2019). Social aversive generalization learning sharpens the tuning of visuocortical neurons to facial identity cues. Technical report, PsyArXiv.

Van Bergen, R. S., Ma, W. J., Pratte, M. S., and Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, 18(12):1728–1730.

Van Damme, S., Lorenz, J., Eccleston, C., Koster, E. H., De Clercq, A., and Crombez, G. (2004). Fear-conditioned cues of impending pain facilitate attentional engagement. *Neurophysiologie Clinique/Clinical Neurophysiology*, 34(1):33–39.

Van Gulick, A. E. and Gauthier, I. (2014). The perceptual effects of learning object categories that predict perceptual goals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5):1307–1320.

Watson, A. B. (2017). Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*, 17(3):10.

Weinberger, N. M. (2007). Associative representational plasticity in the auditory cortex: A synthesis of two disciplines. *Learning & Memory*, 14(1-2):1–16.

Weinberger, N. M. (2011). Reconceptualizing the primary auditory cortex: learning, memory and specific plasticity. In Winer, J. A. and Schreiner, C. E., editors, *The Auditory Cortex*, pages 465–491. Springer US, Boston, MA.

Wolff, M. J. and Rademaker, R. L. (2024). Model mimicry limits conclusions about neural tuning and can mistakenly imply unlikely priors. *bioRxiv*.

Xie, X. Y. and Yu, C. (2018). Double training downshifts the threshold vs. noise contrast (TvC) functions with perceptual learning and transfer. *Vision Research*, 152:3–9.

Yuan, M., Gimenez-Fernandez, T., Mendez-Bertolo, C., and Moratti, S. (2018). Ultrafast cortical gain adaptation in the human brain by trial-to-trial changes of associative strength in fear learning. *Journal of Neuroscience*, 38(38):8262–8276.

Zhang, J., Meeson, A., Welchman, A. E., and Kourtzi, Z. (2010). Learning Alters the Tuning of Functional Magnetic Resonance Imaging Patterns for Visual Forms. *Journal of Neuroscience*, 30(42):14127–14133.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44(1):41–61.