

Novel Tools and Methods

The Importance of Considering Model Choices When Interpreting Results in Computational Neuroimaging

 Thomas C. Sprague,¹  Geoffrey M. Boynton,² and  John T. Serences^{3,4,5}

<https://doi.org/10.1523/ENEURO.0196-19.2019>

¹Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106-9660, ²Department of Psychology, University of Washington, Seattle, WA 98195-1525, ³Department of Psychology, University of California San Diego, La Jolla, CA 92093-0109, ⁴Neurosciences Graduate Program, University of California San Diego, La Jolla, CA 92093-0109, and ⁵Kavli Foundation for the Brain and Mind, University of California San Diego, La Jolla, CA 92093-0126

Abstract

Model-based analyses open exciting opportunities for understanding neural information processing. In a commentary published in *eNeuro*, Gardner and Liu (2019) discuss the role of model specification in interpreting results derived from complex models of neural data. As a case study, they suggest that one such analysis, the inverted encoding model (IEM), should not be used to assay properties of “stimulus representations” because the ability to apply linear transformations at various stages of the analysis procedure renders results “arbitrary.” Here, we argue that the specification of all models is arbitrary to the extent that an experimenter makes choices based on current knowledge of the model system. However, the results derived from any given model, such as the reconstructed channel response profiles obtained from an IEM analysis, are uniquely defined and are arbitrary only in the sense that changes in the model can predictably change results. IEM-based channel response profiles should therefore not be considered arbitrary when the model is clearly specified and guided by our best understanding of neural population representations in the brain regions being analyzed. Intuitions derived from this case study are important to consider when interpreting results from all model-based analyses, which are similarly contingent upon the specification of the models used.

Key words: computational neuroimaging; inverted encoding model; multivariate analysis; stimulus reconstruction

Significance Statement

Gardner and Liu (2019) point out that linear models can provide equally good fits to data across a class of linear transforms applied during analysis. They suggest this is particularly problematic for one analysis method, the inverted encoding model (IEM), that uses activation patterns to estimate responses in modeled information channels, as this renders results arbitrary. Instead, we argue results are not arbitrary when considered in the context of a well-motivated model. Of course, changing model properties can change results, but this applies to all model-based analyses, regardless of inversion. Changing properties of models to recover desired results without disclosure is always ill-advised. When used properly, especially to compare population-level response profiles across conditions, these approaches remain useful tools.

Introduction

Scientists engaging in any research endeavor must necessarily make choices as they perform their research. One can select to measure scalp potentials with EEG or hemodynamic signals with fMRI; one can focus on neural responses in visual cortex or parietal cortex; and one can choose a particular set of modeling assumptions when analyzing data. In some cases, results robustly generalize across such choices. However, it is often the case that the choice made by the researcher matters when interpreting results. For example, scalp potentials index complementary types of attentional modulations to hemodynamic signals (Itthipuripat et al., 2019a), which can impact the conclusions of a study.

When using modern computational models to assay neural function, the modeling choices made by the researchers critically influence results of the modeling procedures. Therefore, it is impossible to interpret the results without knowledge of the details of the model used to generate those results. Moreover, altering the properties of the model should naturally change aspects of the results, sometimes in a predictable and straightforward way. This is true for all models, including the popular single-voxel population receptive field (vRF) modeling approach (Dumoulin and Wandell, 2008; Wandell and Winawer, 2015; Vo et al., 2017), fitting extremely high-dimensional voxel-wise encoding models to densely-sampled datasets (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Huth et al., 2012, 2016; Çukur et al., 2013; Lescroart and Gallant, 2019), the inverted encoding model (IEM) technique (Brouwer and Heeger, 2009, 2011, 2013; Scolari et al., 2012; Foster et al., 2016), and even fitting standard general linear models (GLMs) to task-based fMRI data (Friston et al., 1994).

In a recent commentary, Gardner and Liu (2019) focus on how choices about model specification are meaningful in the context of one of these techniques, the IEM. Typically, this technique involves experimenters estimating the parameters of a simplified model built of stimulus-selective feature channels (e.g., for orientation; color; motion direction; spatial position; polar angle), each tuned to specific feature values and tiling the full stimulus space (Freeman and Adelson, 1991). The properties of these

channels are often inspired by our understanding of the visual system. There are populations of cells tuned to particular orientations; colors; motion directions; positions, and at least in early sensory areas, much is known about the characteristic shape of single-unit tuning functions, and how similarly tuned neurons are clustered along the cortical surface. As an example, one could build a model with eight channels tuned to different stimulus orientations, with each channel modeled with an orientation-selective circular Gaussian tuning function (Brouwer and Heeger, 2011; Ho et al., 2012; Scolari et al., 2012). Using these modeled channels, any stimulus generated from the modeled feature space can be described by the activation of the modeled channels. Then, based on the predicted responses of these modeled channels, linear regression is used to estimate how such a model accounts for changes in activation in each measured signal dimension, typically fMRI voxel or EEG electrode, across different stimulus conditions (fitting the “forward” model). The best-fit model can then be inverted to infer the activation of each modeled channel, that is, the reconstructed channel response profile, given the previously-estimated model and new measured activation patterns across many signal dimensions. The result, when channels are modeled as selective for a single stimulus value, is a channel response profile that typically exhibits a peaked response at the feature value(s) present in the stimulus. Importantly, the inversion step effectively summarizes the results by transforming modulations across all measured signal dimensions (e.g., all voxels or EEG electrodes) back into the model space.

While channel response profiles often look qualitatively similar to neural tuning functions for single units, a point brought up by Gardner and Liu (2019), it is critical to understand that reconstructed channel response profiles cannot be used to draw conclusive inferences about any specific attributes of single-unit response properties (e.g., single tuning width; for more on this, see Sprague et al., 2018a). Moreover, recovery of peaked channel response profiles does not demonstrate the accuracy (or inaccuracy) of the channel shapes in the encoding model used (Sprague et al., 2018a; Gardner and Liu, 2019). We note that Gardner and Liu (2019) use “channel response function” in their commentary, and others have used “channel tuning function,” to refer to results from the IEM technique; we elect to instead use “channel response profiles,” to further distance these results from single-neuron tuning functions.

In their commentary, Gardner and Liu (2019) argue that the channel response profiles resulting from the IEM technique are “arbitrary” because invertible linear transforms of the basis set will fit the data equally well. Hence, changing the shape of the modeled channels can predictably change the shape of the reconstructed channel response profiles. This ability to apply invertible linear transforms means that any reported channel response profile's shape is one from an infinite family of shapes (spanned by all invertible linear transforms that could be applied to the analysis). In their words, “the channel response function is only determined up to an invertible

Received May 24, 2019; accepted November 18, 2019; First published November 26, 2019.

The authors declare no competing financial interests.

Author contributions: T.C.S., G.M.B., and J.T.S. designed research; T.C.S., G.M.B., and J.T.S. performed research; T.C.S. and J.T.S. analyzed data; T.C.S., G.M.B., and J.T.S. wrote the paper.

This work was supported by a Sloan Research Fellowship (T.C.S.) and by National Institutes of Health Grants R01-EY025872 (to J.T.S.) and R01-EY012925 (to G.M.B.).

Acknowledgements: We thank Kirsten Adam, Edward Awh, Clayton Curtis, Joshua Foster, Barry Giesbrecht, Margaret Henderson, and Bradley Postle for helpful discussions and comments on this manuscript.

Correspondence should be addressed to Thomas C. Sprague at tsprague@ucsb.edu or John T. Serences at jserences@ucsd.edu.

<https://doi.org/10.1523/ENEURO.0196-19.2019>

Copyright © 2019 Sprague et al.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

linear transform. Thus, these channel response functions are arbitrary, one of an infinite family and therefore not a unique description of population representation.” (Gardner and Liu, 2019, their abstract). Thus, if a researcher used an unprincipled set of assumptions about the shape of the modeled channels (i.e. ignoring known properties of neural selectivity), then these assumptions can be recapitulated in the reconstructed channel response profiles. For example, Gardner and Liu (2019) showed that if orientation channels are presumed to be bimodal then the resulting reconstructed channel response profiles can also have a bimodal shape.

Below we argue that all models are arbitrary, even those informed by biology, but the results derived from the model are not arbitrary once the model has been specified. This is true for the IEM, but also other neural modeling approaches. Next, we show that even if poorly motivated models are used (or, equivalently, poorly motivated linear transforms are applied), differences between conditions assayed with the IEM technique can be preserved. Finally, we discuss important considerations when interpreting IEM-based analyses and what we see as the place for this modeling approach in the context of other useful analysis methods.

IEM-Based Channel Response Profiles Are Uniquely Determined Given a Fixed Model

It is an unfortunate mischaracterization to imply that IEM-based results are “arbitrary” without specifying that they are uniquely determined and interpretable with knowledge of the encoding model basis used for analysis. Although one can generate many descriptions of a population representation, the result is not arbitrary if the channel response profile is interpreted in the context of the model used by the researchers. As a simple example, one invertible linear transform that could be applied to an encoding model basis and the resulting channel response profiles would shift the columns of the predicted channel response matrix by one. This would result in each channel being mislabeled, but all other features of the analysis would proceed intact. With knowledge of this mislabeling (that is, knowledge of the original basis and the invertible linear transform), it is possible to undo the transform and to achieve the intended understanding. Likewise, if the experimenter reports their basis (as all IEM reports do, so far as we know), and the reconstructed channel responses or derived measures are computed in the context of that basis, then there are no concerns as to the arbitrariness of the channel response profile’s shape. Thus, when principled model basis functions are chosen, it is appropriate to interpret the channel response profile as one possible, but not an arbitrary, depiction of the population representation, as uniquely derived given the principled model choices. That is, IEM results should not be interpreted as revealing *the* population representation; instead, they show one possible depiction of a population representation based on the particular model used.

Results from All Models Depend on Properties of the Model

Importantly, the points Gardner and Liu (2019) raise about applying invertible linear transforms (that is, changing the coordinate system of a linear model) apply to nearly all model-based analyses, even those that only compute a forward encoding model to predict responses of measured neural signals based on stimulus properties, without any attempt at “inversion” back into a stimulus-referred space. We consider two trivial examples: spatial RFs measured via single-unit electrophysiology, and a GLM fit to a two-condition fMRI experiment.

When estimating the spatial RF of a neural measurement (either neuron or voxel), it is necessary to relate the observed neural response to changes in the stimulus. Under certain noise assumptions, one could even weight the stimulus aperture (in screen coordinates) by the observed neural signal. But even this procedure involves an implicit set of model assumptions, namely, that the basis for the stimulus model is in visual field coordinates (one number for each location in the visual field). Thus, the same logic of coordinate transforms applies here: one could apply any number of invertible linear transforms to the image basis and to the estimated RF profile, and the resulting model would account for the same amount of variance because it is a linear transform of the original model. For instance, a 2D Fourier transform could be used to losslessly transform between a spatial basis and a Fourier basis. Does this mean we should consider RF (or feature tuning) models as arbitrary? Of course not. The existence of a potential coordinate transform does not render the original model invalid, it just means that one must know the model to interpret the results.

A similar logic applies to a simple two-condition fMRI experiment using univariate statistical approaches (i.e., voxel-wise analysis with a GLM; Friston et al., 1994). Consider the case where a participant is sometimes pressing a button with their left hand and sometimes a button with their right hand (or looking at pictures of faces or houses, or any other experimental manipulation). The experimenter can build a GLM with predictors for BOLD activation associated with pressing a button with the left and right hand, appropriately convolved with a model hemodynamic response function. In turn, the experimenter could apply the invertible linear transform $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ to the model basis (and thus, the resulting GLM regressors), which would result in flipped estimated β weights: the β weight originally corresponding to right now corresponds to left, and vice versa. However, because you know the original layout of the regressors, you could update your labels of the weights accordingly. While the ability to perform this coordinate transform in principle means the resulting β weights are arbitrarily defined, they remain uniquely and informatively defined given an understanding of the original model. This fact should not be used to label model-based estimates as arbitrary, but instead emphasizes the importance of understanding the model used to derive conclusions about a dataset.

Moreover, it is often the case that as understanding progresses, decisions about how to specify models change. This is evident among researchers developing and fitting voxel RF models: the original, classical demonstration that RF profiles can be fit for individual fMRI voxels implemented a straightforward circular Gaussian encoding model with an assumption of linear scaling (Dumoulin and Wandell, 2008). In subsequent studies, these models were extended to incorporate inhibitory surrounds (Zuiderbaan et al., 2012), compressive spatial nonlinearities (Kay et al., 2013; Mackey et al., 2017), and noncircular RF profiles (Silson et al., 2018), among many others. Does the introduction of a new model invalidate results using simpler models? Typically, it does not. While it is certainly the case that a more accurate model is always preferable, sometimes a simple model can still be useful (e.g., for defining boundaries of retinotopic ROIs). As always, the parameters reported are interpretable given understanding of the model used to derive those parameters.

Differences between Conditions Are Preserved across Linear Transforms of the Basis

At a high level, the IEM technique is a form of model-based dimensionality reduction. This approach estimates a transform from idiosyncratic measurement space (e.g., activation in voxels in V1; α power at EEG scalp electrodes) into a principled, manipulable, model-based “information” space (activation across modeled information channels). Perhaps most importantly, many studies using IEMs seek to compare channel response profiles, or basis-weighted image reconstructions, across task conditions or timepoints in a trial. As described by Sprague et al. (2018a), these studies employ a fixed encoding model, such that activation patterns from different conditions are transformed into the same modeled information space, using a single common estimated encoding model (and often that encoding model is estimated using data from a completely different training task; Sprague et al., 2014, 2016, 2018b; Itthipuripat et al., 2019a,b). In this case, the criticisms raised by Liu et al. (2018) and Gardner and Liu (2019) do not apply: any arbitrary linear transforms would be applied equivalently to the results from each condition; and differences between conditions would be transformed from participant- and stimulus-specific measurement space into the same model-based information space. Invertible transforms would serve only to adjust the axes of the modeled information space, providing a different “view” of the same data. (Note that there may be cases where a transform renders differences between conditions invisible, but this would be exceedingly rare in cases where stimulus features span a feature space.)

To make more concrete the point that differences between conditions can be preserved across linear transforms of the basis, we simulated an fMRI dataset for an experiment that contained two conditions, with one condition evoking a multiplicatively-larger response at the underlying neural level than the other (e.g., an increase in contrast, as in Liu et al., 2018; code available at https://github.com/tommysprague/iem_sim). Briefly, the response

of each of 100 simulated voxels was computed as the sum of the responses of simulated neurons within each voxel, with each simulated neuron having a circular Gaussian tuning function across the feature space (with pseudo-randomly determined tuning bandwidth and amplitude; Fig. 1A). The activity of the neurons within each voxel was computed in response to a set of 8 stimulus orientations across two experimental conditions, with multiplicative gain applied to the simulated neural responses in condition 2 compared to condition 1. One-half of the data, balanced across stimulus type and experimental condition, were designated as a training set and the other half of the data were designated as a testing set. Using data in the training set, we next fit the voxel-wise forward encoding model comprised of eight basis functions that span the feature space using either a standard set of raised cosine basis functions, tuned to specific feature values spanning the orientation space, or a set of raised cosine basis functions that were linearly transformed via an appropriately designed matrix into bimodal basis functions [termed the “xform” matrix (Fig. 1B); mirroring Gardner and Liu (2019)’s Fig. 2; the P matrix in their notation]. We then inverted both forward models, and used those IEMs to reconstruct channel response profiles from the same held-out test data.

Within each condition, channel response profiles recovered a scaled version of the basis function used to estimate the corresponding model (Fig. 1C; mirroring Gardner and Liu (2019)’s Fig. 3]. However, although the shape of the channel response profiles is constrained predictably by the choice of the basis functions, differences between conditions are preserved: condition 2 shows larger-amplitude channel response profiles regardless of the basis used. Importantly, because the transformation is linear and invertible, the bimodal channel response profiles from each condition can be losslessly converted back into unimodal channel response profiles via multiplication with the inverse of the original transformation matrix (Fig. 1C; and note that this holds across a variety of gain modulations and with noise added at the level of simulated neurons, Fig. 1D). Thus, one can apply arbitrary linear transforms to the basis set, and rather than rendering the data arbitrary, they remain interpretable given knowledge of the encoding model.

As shown in Figure 1, although the shape of the channel response profiles is different due to the application of an invertible linear transform, the difference between conditions is preserved. This follows from the fact that, because the end result of the IEM procedure is a linear mapping from signal space into channel space, some differences in measured signals can be detected even across arbitrary basis transforms.

Thus, if the goal is to determine whether the amplitude of the channel response profiles increased, then the application of an invertible linear transform should not impact the general conclusions. Of course, this is true so long as one can accurately quantify or parameterize the resulting shape of the channel response profiles, which may be difficult if a random or oddly-shaped basis is used. Similarly, the process of aligning or re-centering

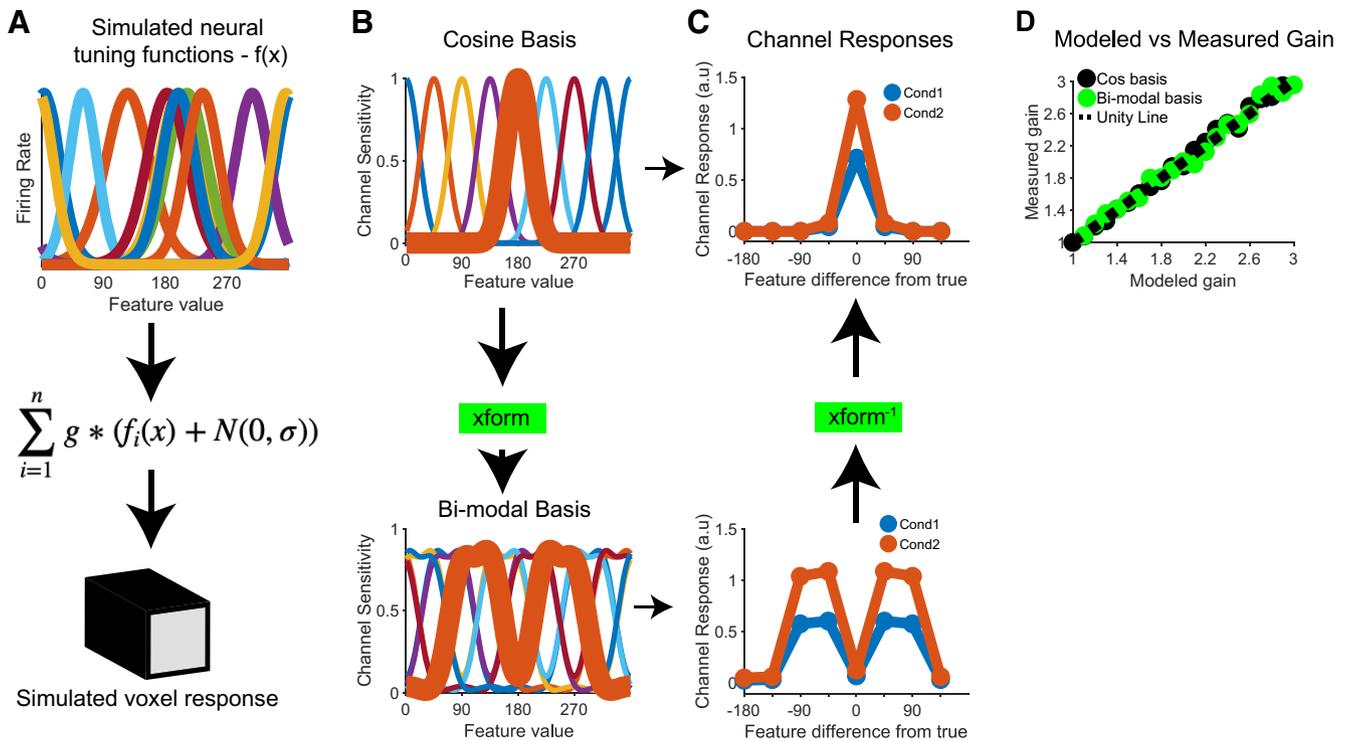


Figure 1. Differences between conditions can be preserved across invertible linear transforms. **A**, We simulated voxel-level fMRI data where each voxel's response was generated based on the sum of simulated responses across a population of simulated neurons with randomly centered tuning preferences and variable bandwidth (here, n = number of neurons, set to 100, although only 10 neural tuning functions are shown for clarity; see code on GitHub for full set of model parameters; https://github.com/tommysprague/iem_sim). Noise was added to the neural responses and then the gain factor (g) was applied to the data from each condition (condition 1: $g = 1$, condition 2: $g = 1.8$). For display purposes the noise (N) was set to 0 for panels **A–C** (following Gardner and Liu, 2019, their Fig. 3) and was set to 10 for panel **D**. **B**, We analyzed data using two different formats of channel basis functions, mirroring those used by Gardner and Liu (2019). Importantly, the two bases are related by an invertible linear transform ($xform$). **C**, Reconstructed channel response profiles differ in similar ways: condition 2 has a higher amplitude than condition 1, regardless of the basis set used, and the bimodal channel response profiles are related by the inverse of the linear transform that was used to create the bimodal basis in the first place ($xform^{-1}$). **D**, Modeled gain compared to measured gain between conditions 2 and 1, computed using both the raised cosine basis set and the transformed bimodal version of the cosine basis set. Because there is not a straightforward way to quantify amplitude for the channel response profiles computed from the bimodal basis, we instead implemented a model-free quantification scheme in which we computed the ratio of the area under each channel response profile (i.e., ratio of area under the curve in condition 2 compared to condition 1).

channel response profiles on the correct feature can become vaguely defined if poorly motivated basis functions are used: typically, a unimodal channel is centered at the feature value to which it is tuned; but a bimodal or other oddly-shaped channel cannot be easily related to a particular feature value, further rendering data presentation and interpretation tricky in such cases. But again, we emphasize that, when channel response profiles are interpreted within the context of the model used to compute them, there is no sense in which the reported result is arbitrary.

Appropriately Selecting Neutral Training Sets

It is not always trivial to decide what type of data to use for estimating a fixed encoding model. Data SNR will likely vary across conditions, and fundamental properties of the encoding model itself may also vary (e.g., spatial RF properties; Vo et al., 2017). Because the IEM can be considered a form of model-based dimensionality reduc-

tion, it is quite important to estimate the most robust model possible. In general, using a neutral “mapping” dataset in which data SNR is optimized is, we believe, the most ideal approach, and can provide a stable “baseline” space in which to transform measured data from an experimental task. For example, when examining how working memory representations change across various task manipulations, we have used a variety of attended visual stimuli for model estimation (Sprague et al., 2016; Rademaker et al., 2019). This is analogous to common procedures used when identifying retinotopic maps on the cortical surface, where researchers typically employ high-contrast flickering checkerboard bars or wedges, and often ask participants to attend the visual stimulus in service of improving SNR (Bressler and Silver, 2010).

It is not always possible to acquire such mapping task datasets, either for experimental or scan time considerations. In this case, there are several possibilities. If multiple conditions are tested (e.g., conditions 1 and 2), one could (1) estimate an encoding model using data from

both conditions 1 and 2 (equal numbers of trials), then use that encoding model to reconstruct held-out data from each condition (Sprague and Serences, 2013), or (2) estimate the encoding model with data from each condition in turn, then reconstruct data from both conditions 1 and 2 using each estimated encoding model. The key feature of such an analysis is that channel response profiles are compared based on a fixed encoding model. If different encoding models are used to reconstruct channel response profiles from different conditions, they may not be on equal footing (Liu et al., 2018), and it can be challenging to determine whether the best-fit encoding model changed, the structure of the neural representation changed, or both changed (Sprague et al., 2018a). In general, using a balanced dataset or the highest-SNR condition for model estimation should likely yield the most robust results, as these datasets will allow for accurate model estimation.

Interpreting Channel Response Profiles

The points raised by Gardner and Liu (2019) offer an opportunity to clarify what can and what cannot be inferred based on results from IEM analyses. In general, the structure of the analysis constrains what can be concluded. We discuss two scenarios: reconstructing channel response profiles within-dataset and using a fixed encoding model.

Computing channel response profiles within-dataset

Often, encoding models are estimated using a subset of runs or trials from an experiment, then the best-fit encoding model is used to reconstruct channel response profiles using the held-out runs or trials. In all other respects, the runs/trials used for model estimation and channel response profile reconstruction are identical. This is often called “leave-one-run-out” or “leave-one-trial-out,” or more generally, *k*-fold cross-validation. In such an analysis structure, the only interpretable result is the positive one: a reconstructed channel response profile peaked over the actual feature value (assuming a sensible model is used). If such a peak is observed after implementing appropriate statistical tests in which null models are used to compute channel response profiles, this offers evidence that the modeled feature value is represented within the measured neural signal space (i.e., ROI; EEG frequency band). That is, if one observes a smooth channel response profile in a leave-one-run-out analysis estimated using a smooth encoding model, this cannot and should not be taken as evidence for a smooth, graded representation of the modeled feature value within the analyzed data (Gardner and Liu, 2019).

However, there is one case where the shape of the channel response profile can be meaningful: if an orthogonal delta or “stick” basis is used, in which no smoothness is imposed on the encoding model, a graded channel response profile is consistent with a smooth representational structure, such that nearby stimulus feature values are represented more similarly in neural signals (Garcia et al., 2013; Saproo and Serences, 2014; Ester et al., 2015, 2016, Foster et al., 2016, 2017; Sutterer et al., 2019). This approximates a form of representational sim-

ilarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013). Note that this does not provide evidence for any particular single-unit tuning properties, but does support a smooth population-level representation as assayed with aggregate neural signals. Finally, in some circumstances, the delta basis can be considered a linear transform of any graded and overlapping basis set typically used, and so the model equivalence issues raised by Gardner and Liu (2019) indeed apply. But, this ability to transform between coordinate systems does not negate an observation of smooth channel response profiles estimated in the context of an orthogonal basis.

Computing channel response profiles with a fixed encoding model

When it is feasible to acquire a unique model-estimation dataset that is only used for the purposes of fitting the encoding model, it is possible to make further conclusions based on the shape of reconstructed channel response profiles. In this analysis structure, the model estimation dataset is never used for reconstructing channel response profiles. Rather, separate “task” data in which one or more experimental conditions are manipulated is used for reconstruction. There is not necessarily any imposed structure on the resulting channel response profiles, and so their properties can be quantified and compared on a fair footing: the IEM fit using separate data offers a static perspective with which to view the measured activation patterns in the task data.

In the context of a fixed encoding model, the observation of peaked channel response profiles in the task dataset suggests that the information is represented in an analogous structure to its representation in the mapping dataset. For example, Rademaker et al. (2019) recently demonstrated that an encoding model estimated with an attentional mapping task (participants monitored an oriented grating for occasional contrast changes) could accurately recover the contents of visual working memory (participants remembered a briefly-presented oriented grating over an extended delay interval) based on activation patterns in occipital cortex.

Changes in channel response profiles between conditions

Under a fixed encoding model, observing different channel response profiles between task conditions offers meaningful insight into the information content of population-level neural representations. Returning to the point that IEMs act as a form of model-based dimensionality reduction, the observation that channel response profiles change across task conditions suggests that the activation pattern is impacted by the experimental manipulation within the particular modeled information space. For example, manipulations of spatial attention are often found to impact the gain of channel response profiles (Garcia et al., 2013; Sprague and Serences, 2013; Itthipuripat et al., 2019a). Due to the linearity of the analysis, this is consistent with a neural mechanism whereby neural populations increase their gain with attention (Kim et al., 2007; Lee and Maunsell, 2010; Fig. 1). However, other mecha-

nisms remain possible. As just one example, attention can impact the spatial RFs of neurons or voxels tuned to nearby locations (Womelsdorf et al., 2006), which can also contribute to an observed change in channel response profile amplitude (Sprague and Serences, 2013; Vo et al., 2017).

Changes in the baseline, width, or even center of channel response profiles can also be meaningfully compared across conditions under a fixed encoding model. For example, Ester et al. (2019) observed biased centers of channel response profiles following category learning, suggesting that human visual cortex represents similar feature values more distinctly when they belong to different learned abstract categories. Changes in the width of channel response profiles are more challenging to interpret, and less commonly observed, but these could be consistent with a uniform change in the tuning width of constituent neural signals (e.g., simulations in Liu et al., 2018). However, these changes are also consistent with other mechanisms, including asymmetric gain across feature channels (Scolari et al., 2012).

While it remains impossible to conclusively determine which mechanism(s) at the single neuron level support an observed change in channel response profiles, they can be compared when examined on equal footing using a fixed encoding model. Importantly, similar interpretational issues arise when considering changes in voxel RF parameters across task conditions (Klein et al., 2014; Kay et al., 2015; Vo et al., 2017), and changes in decoded neural uncertainty (Liu et al., 2018). Many types of changes at the unit level are consistent with observations at the level of single voxels and across large populations of voxels.

What feature is represented?

Additionally, it is important to remember that identification of peaked channel response profiles (or, for that matter, successful decoding of stimulus value) does not unambiguously demonstrate that the modeled feature is represented by the brain signals measured. If a feature value could equivalently be written as a function of another feature, it remains possible that the analysis is sensitive to such confounding signals. For example, there have been several demonstrations that visual orientation (Freeman et al., 2011) and motion direction (Wang et al., 2014) are represented at a coarse scale in human visual cortex, such that the retinotopic position preference of a voxel determines its feature preference. In this case, it is possible that successful reconstruction or decoding of orientation or motion direction is a consequence of these confounded coarse retinotopic signals, rather than fine-grained feature-selective biases within individual voxels. In many cases this isomorphic feature mapping may not matter as experimenters simply want to characterize the representation of information encoded about a stimulus (in whatever format is accessible). However, it is always necessary to consider the relationship between different possible feature spaces when interpreting results from any model-based analysis of feature-selective neural response properties.

Comparison of IEM and Bayesian Approaches to Stimulus Decoding

Gardner and Liu (2019) also make several other points. First, they highlight many positive aspects of the Bayesian decoding approach introduced by van Bergen et al. (2015). We agree that van Bergen and colleagues' (van Bergen et al., 2015; van Bergen and Jehee, 2018) use of a forward model combined with a Bayesian readout rule is an innovative and promising technique, and thoughtfully analyzing data in different ways, especially when employing complex models, is always a good idea. In particular, the Bayesian decoding approach can provide complementary information about the uncertainty with which the activation pattern represents a feature value using an independently-estimated noise model, which is especially useful when trying to link trial-by-trial readouts of neural uncertainty with behavioral measures (van Bergen et al., 2015; van Bergen and Jehee, 2019). That said, we note that the Bayesian approach, like all modeling endeavors, is sensitive to choices made by researchers. For example, analogous to how choices made about channel shapes can impact results from linear IEMs, choices made about the reduced noise model implemented in the Bayesian approach can substantially impact the results (van Bergen and Jehee, 2018). Moreover, invertible linear transforms of the basis set, if not accounted for during the estimation of noise covariance, can lead to large changes in decoding accuracy. As always, motivating analysis choices based on our understanding of neural systems, such as observations that noise covariance scales with tuning similarity (van Bergen and Jehee, 2018), will enable the most robust possible conclusions.

While the Bayesian approach offers many advantages, there are scenarios where directly comparing responses of modeled information channels is more informative. For example, Brouwer and Heeger (2011) compared responses at specific channels across contrast and stimulus conditions to evaluate the impact of cross-orientation suppression, and Ho et al. (2012) and Scolari et al. (2012) compared responses in channels tuned nearby the stimulus orientation across task (emphasize speed vs accuracy) and attention (target left vs target right) conditions. These types of analyses require examining the full response profile across all modeled channels, especially those that are not tuned to the presented stimulus value. This is not easily accomplished with typical decoding analyses, Bayesian or otherwise – that generate a point estimate of the most likely stimulus feature (with or without a corresponding estimate of uncertainty). Moreover, when trying to disentangle responses associated with simultaneously presented stimuli, specifying an appropriate model in the Bayesian framework is not always straightforward. It is necessary to explicitly define a concrete forward model for how multiple stimulus features interact, including, potentially, how they jointly impact the structure of correlated noise. This is certainly not a weakness of the Bayesian approach. It is always ideal to explicitly define model assumptions. However, sometimes it can be useful to visualize how activation patterns are altered within a simple and fixed linear model space, in

which case the traditional IEM approach can be more appropriate.

Instantaneous Decoding vs Extended Model Estimation

Here, we have primarily discussed the IEM technique and a Bayesian decoder based on a formally specified encoding model. There are several other techniques that have been applied to understand neural representations, including RSA (Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013), voxel-wise encoding modeling (Kay et al., 2008; Naselaris et al., 2009, 2011; Nishimoto et al., 2011; Huth et al., 2012, 2016; Çukur et al., 2013; Lescroart and Gallant, 2019), and voxel RF modeling (Dumoulin and Wandell, 2008; Kay et al., 2013; Wandell and Winawer, 2015; Vo et al., 2017). These approaches can be broken down based on their goals: do we wish to build a comprehensive model of a stimulus' representational space within a brain region, including the sensitivity profiles of constituent voxels? Or, do we instead prefer using a simplified model to make inferences about "instantaneous" brain states that can be compared across experimental conditions?

RSA, simplified vRF modeling, and voxel-wise modeling all require participants view stimuli spanning an entire stimulus space to estimate a single result (for RSA: a representational dissimilarity matrix; for voxel-wise modeling, the parameters to each voxel's encoding model). For example, you cannot estimate a voxel's RF profile with stimuli presented in a small number of positions (e.g., just one position on the left and one on the right side of the screen). Instead, you must present stimuli that span the entire stimulus space (e.g., the full extent of the screen, from top to bottom and side to side). Thus, to compare results from any of these methods between task conditions, one must exhaustively sample the stimulus space within each task condition independently. Several studies have taken this approach. Çukur et al. (2013) fit voxel-wise encoding models of semantic space to data acquired while participants viewed natural movies and reported the presence of faces or vehicles (in different scan runs). Accordingly, between scan runs, they could compare properties of the best-fit encoding models across voxels, and concluded that the structure of semantic space is warped between conditions. Several studies have manipulated the locus of spatial attention while participants view visual stimuli used to map vRFs. Klein et al. (2014) and Vo et al. (2017) required participants attend to fixed locations while presenting mapping stimuli, and Sprague and Serences (2013), Kay et al. (2015), Sheremata and Silver (2015), and van Es et al. (2018) all estimated voxel RF profiles during scanning runs when participants either attended to or ignored the mapping stimuli.

RSA and voxel-wise encoding models stand in contrast to the IEM, the Bayesian decoding method, and other decoding approaches. Using these methods, once an encoding model or decoder is estimated (on a held-out set of data), it can be applied to any new activation pattern, even on a time point-by-time point basis. One

case in which this can be used is to compare channel response profiles across a large number of conditions, such as the location of spatial attention and stimulus contrast (Itthipuripat et al., 2019a). In this study, it was not necessary to evaluate channel response profiles at many different stimulus positions, so instantaneous estimates for a small number of stimulus positions offered an efficient means to determine the joint impacts of attention and contrast on population-level stimulus representations. Put another way: there is no way to estimate a voxel-wise encoding model, vRF profile, or representational dissimilarity matrix on a single trial. If single-trial analyses are critical for a given research question, applying an independently-estimated encoding model or decoder is necessary. If, instead, a careful assay of the representational geometry and/or encoding properties of a neural signal are important, analysis of best-fit encoding models to extended datasets should be used.

Finally, the two approaches might be integrated: if trying to understand how a poorly-understood feature space is represented on individual trials, it may be best to start by inferring representational geometry with modern versions of techniques like RSA (Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013; Walther et al., 2016; Cai et al., 2019), then using features of the inferred geometry in combination with model-based analysis methods like IEM or a Bayesian decoder applied to a separate dataset to recover trial-by-trial representations of feature information.

Units of Channel Response Profiles

Gardner and Liu (2019) also point out that the units of model-based reconstructions are arbitrary. This is a point that was noted in one of the original papers to use an IEM (Brouwer and Heeger, 2011). We agree that reconstructed channel response levels are in arbitrary units, and we recommend researchers report them as such going forward. This, combined with the use of unit-normalized modeled channels (i.e., those used to predict channel responses when fitting the forward model), will render channel response estimates more comparable across studies. That said, it is essential to note that these units have no impact on the inferences that can be drawn when comparing channel response functions between conditions under a fixed encoding model. Thus, Gardner and Liu (2019)'s concerns about the arbitrary nature of this scale are not particularly germane to the interpretation of such results: one could scale all units by 42 without impacting the difference between conditions. Thus, so long as all model-based reconstructions that are compared head-to-head are on the same initial footing, then the comparisons are valid regardless of the conventions used to label the units of this analysis. However, when different models are trained for different conditions, it is less certain how to interpret differences in reconstructed channel response profiles across conditions: did the best-fit model, fit individually to each condition, change? Did the data used to reconstruct channel response profiles change? Did both change? By holding at least one aspect constant (the model, estimated with a neutral task or in a

balanced fashion across conditions), it is possible to better ascertain how certain properties of neural response patterns change based on stimulus or task conditions as the units can be compared on equal footing (Sprague et al., 2018a).

IEMs, and Other Analyses Applied to Voxel-Based Measurements, Cannot Be Used to Infer Properties of Single-Unit Tuning

Finally, Gardner and Liu (2019) and Liu et al. (2018) imply that one of the goals of the IEM is to make inferences about single neuron response properties. Making inferences about the response properties of single-neurons is not possible using the IEM or any related model that operates at the scale of aggregate neural signals such as voxels, as different types of single-unit modulations can give rise to identical modulations at the level of a voxel (Sprague et al., 2018a). As described above (see Interpreting Channel Response Profiles), divergent types of changes in single-neuron response properties can lead to identical signals at the aggregate scale. Thus, making such inferences is not the goal of the IEM or related measures, including the Bayesian decoding approach of van Bergen et al. (2015). Instead, a fundamentally different approach that likely requires adopting a different measurement/analysis paradigm, such as parallel characterization of response properties measured across different scales (e.g., fMRI BOLD signal and single-unit electrophysiology; Keliris et al., 2019) would be needed to overcome the ill-posed many-single-neurons-to-voxel mapping problem.

Defining Terms

In the spirit of Gardner and Liu (2019)'s and Liu et al. (2018)'s efforts to delineate the appropriate uses of IEMs, we want to more precisely define several terms related to the IEM technique to help clarify future reports. The IEM technique involves estimating an *encoding model* that best accounts for observed voxel activation responses given stimuli that are transformed into a modeled "channel space" (and under the assumption of linearity such that the response of a given voxel is a linear combination of each of several modeled channels). Once an encoding model is estimated separately for each voxel, that encoding model can be *inverted* and used to *reconstruct* channel response profiles given new measured activation patterns across those same voxels. Those activation patterns are often measured in response to some kind of stimulus (either visual, or something attended, or held in working memory), and the resulting reconstructed channel response profiles typically contain representations of the stimulus/stimuli. To be clear, the result is not strictly a "stimulus reconstruction," but a model-based reconstructed channel response profile. As an example, reconstructed channel response profiles for stimulus orientation are not literally an oriented grating. Instead, they describe the activation of modeled channels in response to a given stimulus, and this description is in a stimulus-referred space. Reconstructed channel response profiles can be

used for several purposes, including *decoding* (recovering the most likely feature value(s) represented, and/or, with the use of an appropriate noise model, their uncertainty) and *quantification* (characterizing the shape of the channel response profile, including "width," "amplitude," etc., which should never be confused with the width or amplitude of single-neuron responses). Of course, all quantification of channel response profiles must be considered in concert with the encoding model used, but if a fixed encoding model is used for reconstructing channel response profiles across several experimental conditions, their properties can be compared in the context of the model.

Conclusions

In this reply to Gardner and Liu (2019), we hope to have clarified some mischaracterizations of how the IEM approach is conducted (see also: Sprague et al., 2018a). To be clear, we are not arguing that the IEM or related approaches are not without serious limitations, the model specification is key, as is understanding what inferences can and cannot be supported by the results (Sprague et al., 2015, 2018a). As Gardner and Liu (2019) point out, these limitations are especially important to recognize when modeling signals in feature spaces that are not well understood, such as those for complex shapes or for higher-order cognitive or social functions. In these situations, an IEM may still be able to quantify differences between conditions and could thus be used to make inferences about changes in the information content of population-level response patterns. However, in this context, drawing unambiguous links between the shape of IEM-derived channel response profiles and the properties of population-level neural representations is not appropriate because changes in the model can result in changes in channel response profiles. Instead, we agree with the suggestions of Gardner and Liu (2019) that careful comparison of forward models that are not related by an invertible linear transform is better suited for this purpose (Brouwer and Heeger, 2009; Nishimoto et al., 2011; Lescroart and Gallant, 2019). That said, IEM-based channel response profiles are not arbitrary when the model choice is based on principled assumptions about neural population representations and, more importantly, channel response profiles are uniquely determined given knowledge of the modeled basis, whatever that basis may be. Finally, similar interpretational issues arise across many types of model-based analyses of neural signals, including voxel RF modeling, voxel-wise encoding models, and RSA. For all these approaches, modeling choices have substantial impacts on results, and so all results must always be interpreted in the context of the model(s) used for analysis.

We believe the IEM method is most useful when comparing reconstructed channel response profiles across manipulations of stimulus properties (e.g., contrast) or task conditions (e.g., attention), or combinations thereof (Sprague et al., 2018b) using a fixed encoding model across relevant comparisons (Sprague et al., 2018a). When used this way, the criticisms raised by Gardner and

Liu (2019) have no substantial bearing on the efficacy of the IEM technique for comparing the impact of experimental manipulations on information represented within aggregate measurements of neural activity patterns. In other words, in the same way the answer (42) is only meaningful in the context of the question (Adams, 1979, 1981), results derived from a model are only meaningful in the context of the model used.

References

- Adams D (1979) *The hitchhiker's guide to the galaxy*, Ed 1. New York, NY: Harmony Books.
- Adams D (1981) *The restaurant at the end of the universe*, 1st American ed. New York, NY: Harmony Books.
- Bressler DW, Silver MA (2010) Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. *Neuroimage* 53:526–533.
- Brouwer G, Heeger D (2009) Decoding and reconstructing color from responses in human visual cortex. *J Neurosci* 29:13992–14003.
- Brouwer G, Heeger D (2011) Cross-orientation suppression in human visual cortex. *J Neurophysiol* 106:2108–2119.
- Brouwer GJ, Heeger DJ (2013) Categorical clustering of the neural representation of color. *J Neurosci* 33:15454–15465.
- Cai MB, Schuck NW, Pillow JW, Niv Y (2019) Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLoS Comput Biol* 15:e1006299.
- Çukur T, Nishimoto S, Huth AG, Gallant JL (2013) Attention during natural vision warps semantic representation across the human brain. *Nat Neurosci* 16:763–770.
- Dumoulin S, Wandell B (2008) Population receptive field estimates in human visual cortex. *Neuroimage* 39:647–660.
- Ester EF, Sprague TC, Serences JT (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron* 87:893–905.
- Ester EF, Sutterer DW, Serences JT, Awh E (2016) Feature-selective attentional modulations in human frontoparietal cortex. *J Neurosci* 36:8188–8199.
- Ester EF, Sprague TC, Serences JT (2019) Categorical biases in human visual cortex. *bioRxiv* 170845.
- Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E (2016) The topography of alpha-band activity tracks the content of spatial working memory. *J Neurophysiol* 115:167–17.
- Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E (2017) Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychol Sci* 28:929–941.
- Freeman WT, Adelson EH (1991) The design and use of steerable filters. *IEEE Trans Pattern Anal Mach Intell* 13:891–906.
- Freeman J, Brouwer GJ, Heeger DJ, Merriam EP (2011) Orientation decoding depends on maps, not columns. *J Neurosci* 31:4792–4804.
- Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 2:189–210.
- García J, Srinivasan R, Serences J (2013) Near-real-time feature-selective modulations in human cortex. *Curr Biol* 23:515–522.
- Gardner JL, Liu T (2019) Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *eNeuro* 6:ENEURO.0363-18.2019.
- Ho T, Brown S, van Maanen L, Forstmann BU, Wagenmakers E-J, Serences JT (2012) The optimality of sensory processing during the speed–accuracy tradeoff. *J Neurosci* 32:7992–8003.
- Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76:1210–1224.
- Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–458.
- Itthipuripat S, Sprague TC, Serences JT (2019a) Functional MRI and EEG index complementary attentional modulations. *J Neurosci* 39:6162–6179.
- Itthipuripat S, Vo VA, Sprague TC, Serences JT (2019b) Value-driven attentional capture enhances distractor representations in early visual cortex. *PLoS Biol* 17:e3000186.
- Kay K, Naselaris T, Prenger R, Gallant J (2008) Identifying natural images from human brain activity. *Nature* 452:352–355.
- Kay KN, Winawer J, Mezer A, Wandell BA (2013) Compressive spatial summation in human visual cortex. *J Neurophysiol* 110:481–494.
- Kay KN, Weiner KS, Grill-Spector K (2015) Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr Biol* 25:595–600.
- Keliris GA, Li Q, Papanikolaou A, Logothetis NK, Smirnakis SM (2019) Estimating average single-neuron visual receptive field sizes by fMRI. *Proc Natl Acad Sci USA* 116:6425–6434.
- Kim YJ, Grabowecy M, Paller KA, Muthu K, Suzuki S (2007) Attention induces synchronization-based response gain in steady-state visual evoked potentials. *Nat Neurosci* 10:117–125.
- Klein BP, Harvey BM, Dumoulin SO (2014) Attraction of position preference by spatial attention throughout human visual cortex. *Neuron* 84:227–237.
- Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci* 17:401–412.
- Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60:1126–1141.
- Lee J, Maunsell JHR (2010) The effect of attention on neuronal responses to high and low contrast stimuli. *J Neurophysiol* 104:960–971.
- Lescroart MD, Gallant JL (2019) Human scene-selective areas represent 3D configurations of surfaces. *Neuron* 101:178–192.e7.
- Liu T, Cable D, Gardner JL (2018) Inverted encoding models of human population response conflate noise and neural tuning width. *J Neurosci* 38:398–408.
- Mackey WE, Winawer J, Curtis CE (2017) Visual field map clusters in human frontoparietal cortex. *Elife* 6:e22974.
- Naselaris T, Prenger R, Kay K, Oliver M, Gallant J (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63:902–915.
- Naselaris T, Kay K, Nishimoto S, Gallant J (2011) Encoding and decoding in fMRI. *Neuroimage* 56:400–410.
- Nishimoto S, Vu A, Naselaris T, Benjamini Y, Yu b, Gallant J (2011) Reconstructing visual experiences from brain activity evoked by natural movies. *Curr Biol* 21:1641–1646.
- Rademaker RL, Chunharas C, Serences JT (2019) Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat Neurosci* 22:1336–1344.
- Saproo S, Serences JT (2014) Attention improves transfer of motion information between V1 and MT. *J Neurosci* 34:3586–3596.
- Scolari M, Byers A, Serences JT (2012) Optimal deployment of attentional gain during fine discriminations. *J Neurosci* 32:1–11.
- Sheremata SL, Silver MA (2015) Hemisphere-dependent attentional modulation of human parietal visual field representations. *J Neurosci* 35:508–517.
- Silson EH, Reynolds RC, Kravitz DJ, Baker CI (2018) Differential sampling of visual space in ventral and dorsal early visual cortex. *J Neurosci* 38:2294–2303.
- Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat Neurosci* 16:1879–1887.
- Sprague TC, Ester EF, Serences JT (2014) Reconstructions of information in visual spatial working memory degrade with memory load. *Curr Biol* 24:2174–2180.

- Sprague TC, Saproo S, Serences JT (2015) Visual attention mitigates information loss in small- and large-scale neural codes. *Trends Cogn Sci* 19:215–226.
- Sprague TC, Ester EF, Serences JT (2016) Restoring latent visual working memory representations in human cortex. *Neuron* 91:694–707.
- Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA (2018a) Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *eNeuro* 5:ENEURO.0098-18.2018.
- Sprague TC, Itthipuripat S, Vo VA, Serences JT (2018b) Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. *J Neurophysiol* 119:2153–2165.
- Sutterer DW, Foster JJ, Serences JT, Vogel EK, Awh E (2019) Alpha-band oscillations track the retrieval of precise spatial representations from long-term memory. *J Neurophysiol* 122:539–551.
- van Bergen RS, Jehee JFM (2018) Modeling correlated noise is necessary to decode uncertainty. *Neuroimage* 180:78–87.
- van Bergen R, Jehee JFM (2019) Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *J Neurosci* 39:8164–8176.
- van Bergen RS, Ma WJ, Pratte MS, Jehee JFM (2015) Sensory uncertainty decoded from visual cortex predicts behavior. *Nat Neurosci* 18:1728–1730.
- van Es DM, Theeuwes J, Knapen T (2018) Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. *Elife* 7:e36928.
- Vo VA, Sprague TC, Serences JT (2017) Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. *J Neurosci* 37:3386–3401.
- Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* 137:188–200.
- Wandell BA, Winawer J (2015) Computational neuroimaging and population receptive fields. *Trends Cogn Sci* 19:349–357.
- Wang HX, Merriam EP, Freeman J, Heeger DJ (2014) Motion direction biases and decoding in human visual cortex. *J Neurosci* 34:12601–12615.
- Womelsdorf T, Anton-Erxleben K, Pieper F, Treue S (2006) Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nat Neurosci* 9:1156–1160.
- Zuiderbaan W, Harvey BM, Dumoulin SO (2012) Modeling center-surround configurations in population receptive fields using fMRI. *J Vis* 12:10.