An improved method for evaluating inverted encoding models

Paul S. Scotti, Jiageng Chen, & Julie D. Golomb

Department of Psychology, The Ohio State University, Columbus, Ohio, USA

Please address correspondence to:
Paul S. Scotti
The Ohio State University
Department of Psychology
1835 Neil Avenue
Columbus, OH 43210
Email: scottibrain@gmail.com

Abstract word count: 193
Main text word count: 5110
Figures: 5

Abstract

Inverted encoding models have recently become popular as a method for decoding stimuli and investigating neural representations. Here we present a novel modification to inverted encoding models that improves the flexibility and interpretability of stimulus reconstructions, addresses some key issues inherent in the standard inverted encoding model procedure, and provides trial-by-trial stimulus predictions and goodness-of-fit estimates. The standard inverted encoding model approach estimates channel responses (or “reconstructions”), which are averaged and aligned across trials and then typically evaluated using a metric such as slope, amplitude, etc. We discuss how this standard procedure can produce spurious results and other interpretation issues. Our modifications are not susceptible to these methodological issues and are further advantageous due to our decoding metric taking into account the choice of population-level tuning functions and employing a prediction error-based metric directly comparable across experiments. Our modifications also allow researchers to obtain trial-by-trial confidence estimates independent of prediction error which can be used to threshold reconstructions and increase statistical power. We validate and demonstrate the improved utility of our modified inverted encoding model procedure across three real fMRI datasets, and additionally offer a Python package for easy implementation of our approach.

1. Introduction

A mental representation can be defined as the "systematic relationship between features of the natural world and the activity of neurons in the brain" (Poldrack, 2020). Encoding models describe this relationship computationally, typically by reducing the complexity of the input data with a set of functions that, when combined, roughly approximate the neural signal. Encoding and decoding models (aka voxel-based modeling or stimulus-model based modeling) have become a standard method for investigating neural representational spaces and predicting stimulus-specific information from brain activity (Naselaris, Kay, Nishimoto, & Gallant, 2011; Naselaris & Kay, 2015; Popov, Ostarek, & Tenison, 2018). The key advantages of such models over other computational approaches such as multivariate pattern classification, representational similarity analysis, or population receptive field mapping are typically touted as the following: (1) Encoding models can take inspiration from single-unit physiology by consisting of tuning functions in stimulus space (aka feature space), allowing both the maximally receptive feature and the precision/sensitivity of the tuning to be estimated across a population of neurons; (2) The encoding model that transforms stimuli into brain activity can be inverted into a decoding model capable of predicting stimuli given a pattern of neural activity; and (3) The decoding model can predict novel stimuli or experimental conditions not used in the training of the model. The features of an encoding model can be anything from Gabor filters (Kay et al., 2008; Naselaris et al., 2009) to perceptual colors (Brouwer & Heeger, 2009) to acoustic musical features (Casey et al., 2012) and even human faces (Lee & Kuhl, 2016).

The inverted encoding model (IEM) is one example of an encoding and decoding model that uses simple linear regression and a basis set representing the hypothesized population-level tuning functions, consisting of several channels that are modeled as cosines (or von Mises) equally separated across stimulus space (e.g., orientation, color, spatial location). Due to its simplicity, robust performance, and grounding in single-unit physiology principles, inverted encoding models have quickly risen to prominence in the cognitive neuroscience community (e.g., Ester, Sprague, & Serences, 2020; Foster, Bsales, & Awh, 2020; Henderson, Vo, Chunharas, Sprague, & Serences, 2019; Kim, Hong, Shevell, & Shim, 2020; Kok, Rait, & Turk-Browne, 2020; Lorenc, Vandenbroucke, Nee, de Lange, & D'Esposito, 2020; Oh, Kim, & Kang, 2019; Rademaker, Chunharas, & Serences, 2019; Sutterer, Foster, Adam, Vogel, & Awh, 2019; Yu, Teng, & Postle, 2020). The basic idea behind IEMs is that each channel in the basis set can be assigned a weight per voxel[1] and hence a model can be trained to predict the activity of each voxel using the weights of each channel as predictors (i.e., the regressors in a linear regression). Then, this trained encoder is inverted such that it becomes a decoder capable of predicting, or *reconstructing*, a trial's stimulus when provided with a novel set of voxel activations.

Here we present a novel modification to IEMs that improves the interpretability of stimulus reconstructions, addresses some key issues inherent in the standard IEM procedure, and provides trial-by-trial stimulus predictions and goodness-of-fit estimates.

[1] In this paper we adopt fMRI nomenclature (voxels), but IEMs have been successfully applied to neuroimaging modalities besides fMRI, including EEG and MEG, and our modified procedure can be applied to other modalities as well.

In the sections that follow, we briefly review the standard procedure used for implementing IEMs, present an overview of our modified procedure, and discuss methodological issues and limitations with the standard procedure that our modified procedure addresses. We then present results comparing the standard procedure to our modified procedure using three existing fMRI datasets. These results validate our approach and highlight its practical advantages in terms of improved flexibility and interpretability. We also offer a publicly available Python package for researchers to easily implement inverted encoding models using our modified approach.

1.1 Standard procedure for inverted encoding models

The standard IEM procedure is illustrated in Figure 1A-B using a toy example where a participant was shown eight trials of colored squares and the researcher used an IEM to reconstruct the presented colors based on the pattern of activity in a six-voxel brain region. Note that this “standard procedure” is our depiction of the typical steps used to implement IEMs in the neuroimaging literature, but there may be nuances particular to specific papers that deviate from this procedure.

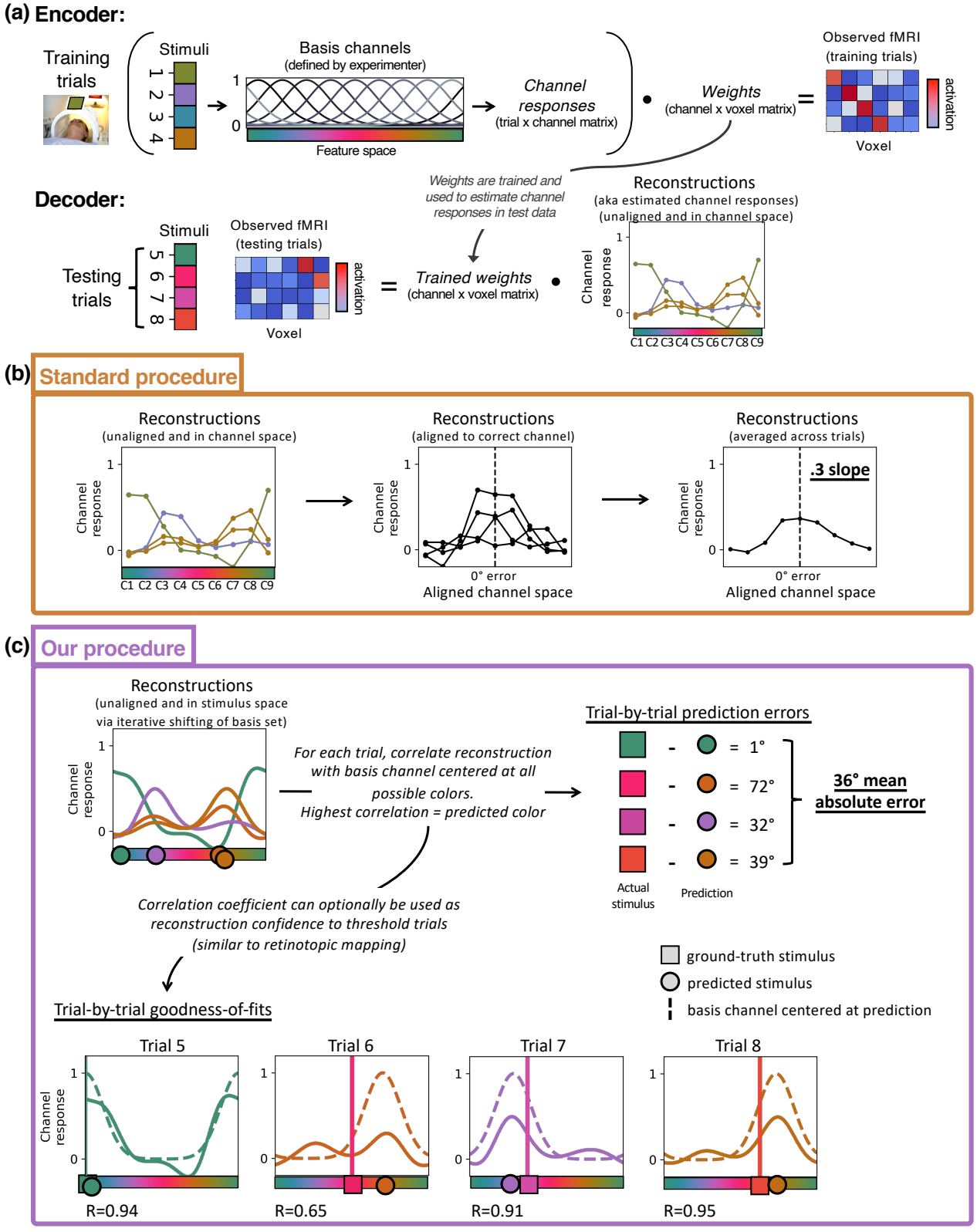

Figure 1. Simulated example of an fMRI experiment using the standard IEM procedure and our modified procedure, where a participant was shown eight trials of colored squares and the

researcher used an IEM to decode the presented colors with a six-voxel brain region. (a) The top row depicts the encoding model where the weights matrix is estimated (via linear regression) and the second row depicts the decoding model where the channel responses for the test data are estimated using the trained weights from the encoding model. (b) The standard procedure involves aligning and averaging reconstructions and measuring the result according to a variety of possible metrics (e.g., amplitude, slope). (c) Our modified procedure deviates from the standard procedure by evaluating prediction errors rather than an averaged reconstruction. We correlate the basis channel with reconstructions to estimate predicted stimuli, use iterative shifting of the basis set to allow channel space to equal stimulus space, and estimate goodness-of-fits for each trial reconstruction which can be used as a measure of confidence for each trial's predicted stimulus. For simplicity, this example shows the encoder trained on the first half of trials and the decoder used to predict the color of the remaining trials, but in most applications cross-validation should be used such that every trial may be decoded while avoiding circularity/double-dipping.

The prerequisites for implementing an IEM are (1) an array specifying the feature values of the presented stimulus for every trial, (2) a trial-by-voxel matrix containing the observed brain activations[2] for every voxel per trial, and (3) a basis set representing the hypothesized population-level tuning functions. Typically, researchers use a basis set consisting of several equidistant channels modeled as cosines raised to the number of channels minus 1. This means that each channel is modeled as

$$\cos\left((\theta-\mu)\frac{\pi}{stimulus\_range}\right)^{num\_channels-1}$$

where θ is degrees in stimulus space, μ is the center of each channel, and *stimulus_range* is the range of stimulus space (e.g., 360° for hues on a color wheel spanning 0-359°). The reasoning behind raising cosines to the *num_channels*-1 is to make the tuning curves narrower and more comparable to physiological findings (Brouwer & Heeger, 2011), and the specification of the number of channels is mostly

[2]Observed brain activations could be beta weights from general linear model estimation (e.g., Mumford, Turner, Ashby, & Poldrack, 2012) or raw BOLD signal from a block or slow-event related design.

arbitrary (more channels are typically chosen if suspecting narrow tuning and vice-versa for broad tuning).

The encoder models each voxel's response as the weighted sum of the channels, such that the observed trial-by-voxel fMRI activation matrix is equal to the dot product of the basis set and the weight matrix,

$$basis\_set[trial\_features,:] \cdot channel\_by\_voxel\_weights = trial\_by\_voxel\_activation$$

where *trial_features* is the feature (e.g., color) of the stimulus and *basis_set* is the matrix of channels with shape (*stimulus_range, num_channels*) described above. Given that the trial-by-voxel matrix and the basis set are already given, the weights matrix can be estimated via least-squares linear regression.

Once the weights matrix is estimated from the training dataset, it can be inverted such that the encoder becomes a decoder for the test dataset. Now, instead of estimating the weights matrix via least-squares linear regression, the weights matrix and the trial-by-voxel matrix are given and the channel responses (i.e., reconstructions) are estimated.

$$trial\_by\_voxel\_activation \cdot channel\_by\_voxel\_weights^{-1} = reconstructions$$

The resulting estimated channel responses, or simply *reconstructions*, is a trial-by-*num_channels* matrix where each trial has its own reconstruction composed of weighted cosines.

In the standard procedure, all the trial-by-trial reconstructions are then circularly shifted along the x-axis such that the channel that *should* have been maximally responsive on every trial (i.e., the channel closest to the ground truth stimulus feature) is aligned to the center of the x-axis (Figure 1B). The aligned reconstructions are then

averaged together to result in a single reconstruction. Assuming that the model performs well, the averaged reconstruction should resemble the shape of the original basis channel with the highest point centered on the aligned location. The averaged and aligned reconstruction is then typically assessed using a number of possible metrics (see Figure 2). As Figure 2 illustrates, there is substantial variability in the choice of metrics used to evaluate IEMs, with the most common metrics being the amplitude or slope of the aligned and averaged reconstruction, or assessing reconstruction quality following a model fitting step.

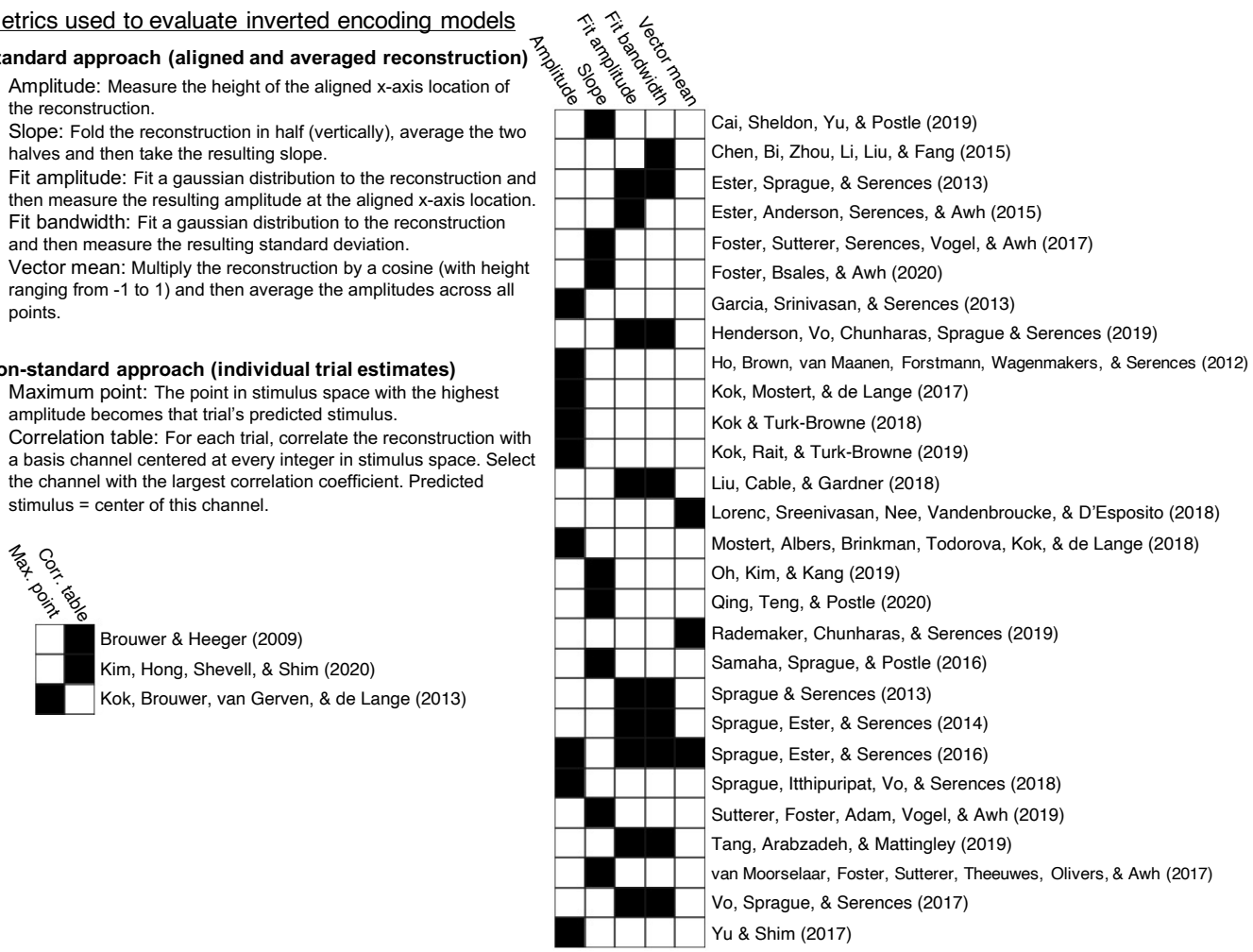

Figure 2. Summary of metrics used to evaluate IEM reconstructions in a sampling of published papers. Note that the methodological concerns about spurious conclusions raised in Section 1.3.1 apply to the metrics labeled under the standard approach, although our proposed modifications pose improvements over typical applications of "maximum point" and "correlation table" approaches as well. The non-standard approaches can be quantified with a single value (like the standard approach metrics) by taking the mean absolute error between predicted and actual stimuli.

Typically, the chosen reconstruction metric is calculated and then subjected to statistical analysis (e.g., permutation testing to assess whether the reconstruction metric is significantly different from what would be expected to occur by chance, or comparisons between reconstruction metrics obtained under different conditions). For permutation testing, the stimulus labels are randomly shuffled, and the average

reconstruction is evaluated for every iteration. These iterations form a null distribution (e.g., 1,000 slopes for 1,000 iterations) and then the actual reconstruction's measure is compared to this null distribution.

1.2 Overview of our modified IEM procedure

The core steps depicted in Figure 1A – using least-squares linear regression for estimating the channel weights (for the encoder) and estimating the channel responses (for the decoder), as well as the use of a basis set of hypothesized population-level tuning functions – remain the same between the standard and our modified IEM procedures. Our modified procedure (Figure 1C) differs from the standard procedure in three key ways. As a brief overview, our first modification is to repeat the entire IEM procedure multiple times with slightly shifted basis sets such that reconstructions are in stimulus space rather than channel space. This iterative shifting modification has been employed in a few previous papers (Kim, Hong, Shevell, & Shim, 2020; Rademaker, Chunharas, & Serences, 2019), however, it is not common practice. This step is important because otherwise stimulus predictions will be biased by the arbitrary placement of the basis channels, as described in section 1.3.2.

Second, and most critically: we evaluate reconstructions in terms of average *prediction error* instead of the aligned and averaged reconstruction metrics described above. Trial-by-trial predictions may be compared to the ground truth stimuli to calculate each trial's prediction error, which may be averaged across trials and assessed via permutation testing just like the standard IEM procedure. Our modified IEM procedure obtains trial-by-trial stimulus predictions using the correlation table approach (the only approach for obtaining a decoding metric in Figure 2 that adapts to the shape of the

basis channel, see section 1.3.1) and calculates prediction error using mean absolute error (MAE), which we recommend as a simple and interpretable metric. (Note that other error metrics are also possible in this framework, including signed error metrics if there is reason to expect asymmetric reconstructions.) This modification is critical because it resolves several methodological concerns inherent in the standard approach, as described more below, and prediction error is an easily interpretable metric.

Our final modification takes the trial-by-trial prediction approach described above one step further. Using the correlation table approach to determine the best-fitting basis channel, the center of that best-fitting channel is taken as the stimulus prediction, but the goodness-of-fit (correlation) values themselves can also be optionally leveraged to estimate trial-by-trial confidence of predictions (see section 1.3.3). This modification adds substantial flexibility to the IEM procedure; e.g., allowing for thresholding reconstructions to potentially increase statistical power, as we demonstrate in the Results. We discuss these modifications in more depth in the subsequent sections, while highlighting the advantages of our modified procedure over the standard procedure.

1.3 Value of our approach over the standard approach

As summarized above, the modified IEM approach that we present here is a combination of several modifications and improvements on the standard IEM procedure. The value of these modifications is primarily in terms of evaluating IEM results: We propose that our modified approach is better than the standard IEM approach in terms of improved interpretability, flexibility, and robustness to methodological concerns. We

also offer our modified approach as a standardized set of “best practices”. As depicted in Figure 2, various approaches exist for evaluating IEMs and often researchers report several decoding metrics due to ambiguity over which metric is best. Our modified approach is the combination of specific practices (some previously employed, some novel) intended to offer a preferred solution.

Below we describe several methodological concerns and limitations of the standard IEM procedure that are addressed by our modified procedure. (In the Results section, we further demonstrate the appeal of our approach in terms of improved flexibility and interpretability.)

1.3.1 Standard procedure can produce misleading or difficult to interpret results

The standard procedure is susceptible to inappropriate decoding evaluations, largely due to the align-and-average step. Aligning and averaging across trial reconstructions loses information that is important for evaluating decoding performance and can be prone to heavy bias from outliers. Moreover, the metrics used to evaluate the aligned and averaged reconstructions are not easily interpretable.

As depicted in Figure 3, averaging can obscure important information present in trial reconstructions. Panels 3a and 3b would be interpreted identically according to the standard procedure even though one example shows every channel correctly predicted (i.e., predicting the correct stimulus feature with minimal error) and the other example shows every channel incorrectly predicted (large errors). Our modified approach using MAE would correctly identify the first case as demonstrating superior decoding performance. The takeaway here is that averaging across prediction errors, and not

across trial reconstructions, avoids the pitfall of interpreting Figures 3a and 3b as reflecting the same level of stimulus-specific brain signal despite clear support for Figure 3a demonstrating improved decoding on a trial-by-trial level.

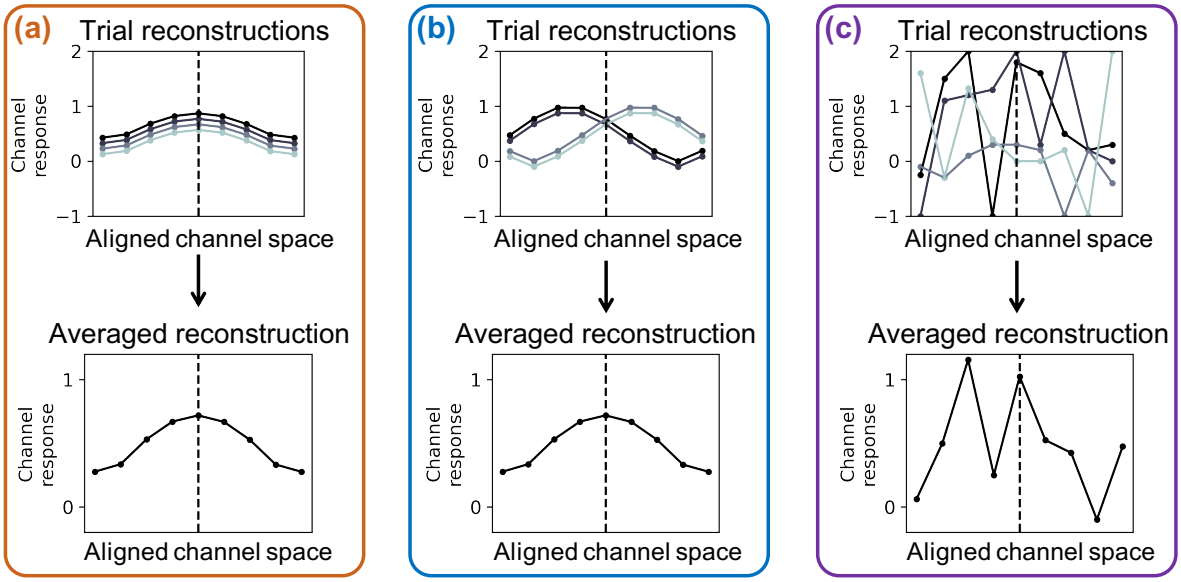

Figure 3. Cartoon example depicts some problems with the standard procedure of evaluating the aligned and averaged reconstruction and using a decoding metric that does not consider the shape of the basis channel or the variability of trial reconstructions. For each of the 3 simulated data examples, the top row depicts four single-trial reconstructions, and the bottom row depicts the aligned-and-averaged reconstruction. In (a) each individual trial's reconstruction accurately predicts the correct channel (i.e., the correct stimulus feature), appropriately reflected in the averaged reconstruction. In (b) each individual trial's reconstruction predicts an incorrect channel. Averaging across trials leads to a misleading result, i.e., the standard approach would consider (b) to reflect the same level of decoding performance as (a). In (c) each individual trial's reconstruction is essentially noise, such that the averaged reconstruction results in a false peak around the aligned point; the standard procedure using align-and-average metrics would result in spuriously superior decoding performance than both (a) and (b), with (c) having a higher amplitude, steeper slope, and narrower standard deviation when fit with a gaussian distribution. Our modified procedure, calculating MAE from trial-wise prediction error, would correctly conclude that case (a) shows the best decoding performance.

MAE is also less prone to bias from outlier reconstructions compared to any of the align-and-average metrics. In the standard procedure, a single outlier reconstruction

can disproportionately bias the averaged reconstruction, potentially completely flipping the averaged reconstruction in the most extreme cases. In contrast, imagine an example worst-case scenario for MAE where an experiment is composed of 300 trials and 299 trials predicted the correct stimulus and one trial predicted the stimulus 180° away (assuming 360° stimulus space). The result would remain at near-perfect performance at 0.6° MAE. For 100 trials the performance would be 1.8° MAE and for 1,000 trials performance would be 0.18° MAE. Outlier bias is minimized because the worst possible prediction error for a single trial is capped at the range of stimulus space divided by two (for circular stimulus spaces), whereas there is no defined limit for the standard IEM procedure.

1.3.2 Standard procedure does not account for the shape of the basis channels

IEMs produce reconstructions that depend on the choice of basis set (e.g., Liu, Cable, & Gardner, 2018; Sprague, Boynton, & Serences, 2019). The decoding metrics commonly used to evaluate averaged reconstructions in the standard IEM procedure, however, do not take this observation into account. That is, intuition – and standard practice – wrongly assume that a monotonic relationship exists between decoding metrics such as slope, amplitude, and bandwidth and a greater amount of stimulus-specific information in the brain signal. A perfect reconstruction returns the shape of the basis channel, and so it makes sense to compare the shape of the reconstruction to the shape of the basis channel to make predictions and evaluations. The correlation table approach employed in our modified procedure leverages this observation to provide the most direct relationship between IEM performance and stimulus-specific brain signal.

The correlation table approach operates as follows and has been previously used in a small number of papers (e.g., Brouwer & Heeger, 2009; Kim, Hong, Shevell, & Shim, 2020). For each trial, compute a set of correlation coefficients, each reflecting the correlation between that trial's reconstruction and a canonical basis channel (i.e., a perfect reconstruction) centered at every integer in stimulus space (e.g., resulting in 360 correlation coefficients for a stimulus space ranging from 0-359°). The highest of these correlation coefficients is determined to be the best fit for that trial, and the predicted stimulus feature for that trial is simply the center of that best-fitting basis channel. In this manner, the predictions obtained from the correlation table metric automatically adjust to consider the shape of the basis channel because it is the basis channel itself that is being used to obtain predictions.

Simply put, amplitude, slope, bandwidth, etc. are inferior metrics compared to the correlation table metric because they do not adapt to the choice of basis set. For instance, using the amplitude metric, a higher amplitude at the aligned point is thought to reflect improved performance. If the basis channel ranges from 0 to 1, a perfect reconstruction should have an amplitude of exactly 1 at the aligned point, but reconstructions can feasibly have amplitudes far greater than 1. Such a problem is demonstrated visually in Figure 3 where it is clear that Figure 3c looks to be a worse reconstruction than Figure 3a, but align-and-average metrics would produce spuriously high values for this case.

It is possible to partially account for the shape of the basis channel by, for example, fitting the reconstruction with a gaussian distribution (e.g., Henderson et al., 2019). However, such fitting procedures may be problematic because such a procedure

forces the reconstruction to appear to be a reasonable gaussian shape regardless of the data (e.g., fitting Figure 3c with a gaussian distribution would still lead to the same incorrect conclusion of superior decoding performance compared to Figure 3a).

Given the various ways to measure IEM performance listed in Figure 2, the correlation table approach best takes the shape of the basis set into account, but one limitation is that the basis channels are in stimulus space, but the reconstructions are in channel space. That is, to properly correlate the shape of the basis channel to the reconstruction, one must linearly interpolate between points in channel space. Our modified procedure solves this limitation by employing iterative shifting (e.g., Kim, Hong, Shevell, & Shim, 2020; Rademaker, Chunharas, & Serences, 2019). By repeatedly fitting the encoding model with every possible (circular) shift of the basis set and then combining all of these iterations together, a fuller reconstruction is obtained that is no longer impoverished by a limited number of *num_channels* points (i.e., the range of channel space becomes equal to the range of stimulus space).

The iterative shifting procedure also aids more generally in producing more interpretable and less biased reconstructions, as illustrated in Figure 4. Iterative shifting of the basis set is especially important because our decoding model must be capable of predicting any possible feature in stimulus space (that is, not solely the stimuli that are located at the centers of the basis channels) if we want to obtain accurate trial-by-trial stimulus predictions. Note that iterative shifting does not change the fact that different basis sets result in different reconstructions, rather, it simply allows for the most accurate reconstruction given a set number of channels with defined bandwidths.

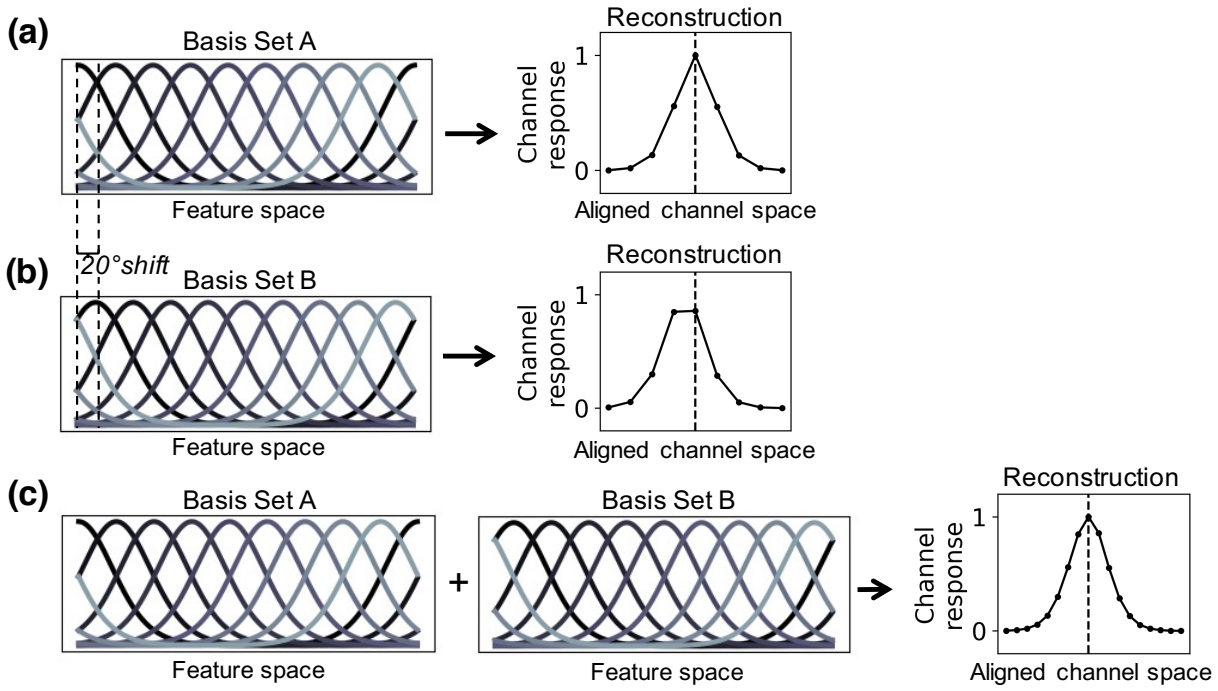

Figure 4. Simulated data depicting how a slightly altered basis set (means shifted by 20°) can alter reconstructions, even if the same signal is present in all cases. Here the trial by voxel activations reflect perfect (zero noise) information with identical train and test sets, such that the resulting reconstructions *should* also be perfect. (a) Basis set perfectly reflects the underlying voxel tuning functions (simulated ground truth). (b) Reconstruction of the same data, now with basis set of channels circularly shifted 20°. (c) By combining the results of both basis sets, the channel space changes from *num_channels* to *num_channels**2, leading to a fuller reconstruction. Iterative shifting in our modified procedure repeats this procedure for all possible shifts of the basis set to make the channel space equal the stimulus space, decreasing variation and allowing the correlation table metric to be optimally applied. The code to reproduce this figure from simulated data can be found at https://osf.io/et7m2/ (also contains code for reproducing Figures 1 and 3).

1.3.3 Standard procedure lacks a measure of decoding uncertainty

Another limitation of the standard approach resolved by our modified approach is that the standard IEM procedure does not incorporate uncertainty into decoding performance. Individual trials can vary substantially in signal quality, driven by factors including attentional fluctuations, alertness, head motion, and scanner noise. Noisier trials could potentially obscure an underlying signal, but as exemplified in Figure 3,

highly variable trial reconstructions are not weighted differently from robust reconstructions according to the standard procedure. The lack of uncertainty information has been noted in other contexts, with some recent alternatives to IEM proposed to incorporate uncertainty (e.g., TAFKAP probabilistic decoding model: Li, Sprague, Yoo, Ma, & Curtis, 2021; van Bergen & Jehee, 2021). However, here we demonstrate that our modified procedure can easily and automatically produce a trial-by-trial measure of prediction uncertainty within the IEM framework itself, which can then be used in flexible and accessible ways.

The correlation table approach produces a best-fitting stimulus prediction – and associated goodness-of-fit value (correlation coefficient) – for each trial. We propose that the correlation coefficient of the best-fitting basis channel can be used as a proxy to estimate the confidence, or reliability, of trial-by-trial predictions. It is important to emphasize that the correlation coefficient reflects the degree to which the reconstruction matches the *best-fitting* basis channel, not the basis channel centered on the correct stimulus. In other words, this goodness-of-fit information is obtained independently and prior to any calculation of prediction error.

This trial-by-trial prediction uncertainty information could be used in a number of different ways. One suggestion we put forth is that goodness-of-fit can be used to threshold reconstructions, such that worse-fitting trials may be excluded from analysis. This principle is analogous to the phase-encoded retinotopic mapping and population receptive field modeling techniques, where a set of models spanning the full stimulus space is evaluated for every voxel, and the parameters of the best-fitting model are selected as that voxel's preferred stimulus, with the goodness-of-fit values then used to

threshold the results (Dumoulin & Wandell, 2008; Engel, Glover, & Wandell, 1997; Sereno et al., 1995).

In the Results section we provide a proof of concept using real fMRI data to demonstrate the utility of using trial-by-trial goodness-of-fit values to threshold IEMs based on confidence. Note that although r-squared is the more commonly used statistic for computing goodness-of-fit in linear regression, squaring the correlation coefficient is not preferred here because the sign of the correlation coefficient is informative (e.g., a perfectly inverted reconstruction should *not* be assigned equal confidence as a perfect reconstruction), so we recommend the use of the r-statistic.

2. Results

To validate our modified IEM procedure and demonstrate its practical advantages, we implemented both the standard IEM procedure and our modified IEM procedure across three real fMRI datasets. The three datasets (Chen, Scotti, Dowd, & Golomb, 2021; Henderson, Vo, Chunharas, Sprague, & Serences, 2019; Rademaker, Chunharas, & Serences, 2019) span the research topics of perception, attention, and memory. We demonstrate how our modified procedure improves over the standard procedure in terms of flexibility (potential to exclude low-confidence trial reconstructions) and interpretability (reconstruction performance in terms of prediction error) while avoiding the methodological pitfalls discussed in the previous section. See the Online Methods for information regarding each dataset and how data were obtained and processed.

2.1 Validating our method on real fMRI data

Figure 5 shows the results from both a standard IEM procedure and our modified procedure. First, to validate our method, we confirmed that across all three datasets our modified IEM replicated the overall pattern of results obtained with the standard IEM. In the Perception dataset (Henderson et al., 2019), we used both techniques to decode the horizontal position of a stimulus in V1, V4, and IPS. The standard IEM procedure (quantified by the slope metric) revealed significant decoding performance in all 3 ROIs, with the strongest decoding (greatest slope) in V1, followed by V4, and then IPS. The modified IEM replicated this pattern. In the Attention dataset (Chen et al., 2021), we used both techniques to decode the attended color within a multi-item, multi-feature stimulus array, in the same three ROIs. The standard IEM procedure revealed significant decoding performance in V1 and V4, but not IPS. The modified procedure again replicated this pattern. Finally, in the Memory dataset (Rademaker, Chunharas, & Serences, 2019), we used both techniques to decode the remembered orientation of a stimulus over two types of working memory delays, blank and with distractor. With the standard procedure, the remembered orientation could be successfully decoded in V1, V4, and IPS, with significantly greater decoding during the blank (vs distractor) delay in V1 and V4. With the modified procedure, we replicated each of those results, with the additional finding of significantly greater decoding during the blank vs distractor delay in IPS as well.

The important takeaways from these validations are that (1) the modified procedure is not susceptible to the methodological concerns raised earlier that plague the standard procedure, and yet (2) the results from the two techniques produce largely

consistent patterns across these datasets, in terms of significance testing and overall patterns of decoding. We note that these three datasets were useful validation cases because they contained robust results (as may be more likely with published, publicly available datasets); however, we would not necessarily expect the modified and standard procedures to always produce consistent patterns, especially in cases where data are less robust and therefore more susceptible to the aforementioned methodological concerns. In those cases, we argue that the modified procedure offers a more accurate reflection of decoding performance, as illustrated in Figure 3.

2.2 Demonstrating the improved flexibility and interpretability of our method

Having validated our method across three diverse fMRI datasets, we next use these same datasets to illustrate the practical advantages of our modified IEM method. First, the modified procedure yields metrics that are easily interpretable and comparable across datasets due to decoding performance being measured in terms of prediction error rather than arbitrary units. For example, in the Memory dataset, the standard procedure in V1 results in decoding performance with a slope of .006 for the blank delay condition and .004 for the distractor delay condition; the modified procedure replicates this pattern, but now with a more interpretable and meaningful metric: orientation can be decoded with an error of 27.9 degrees in the blank delay and 32.5 degrees in the distractor condition.

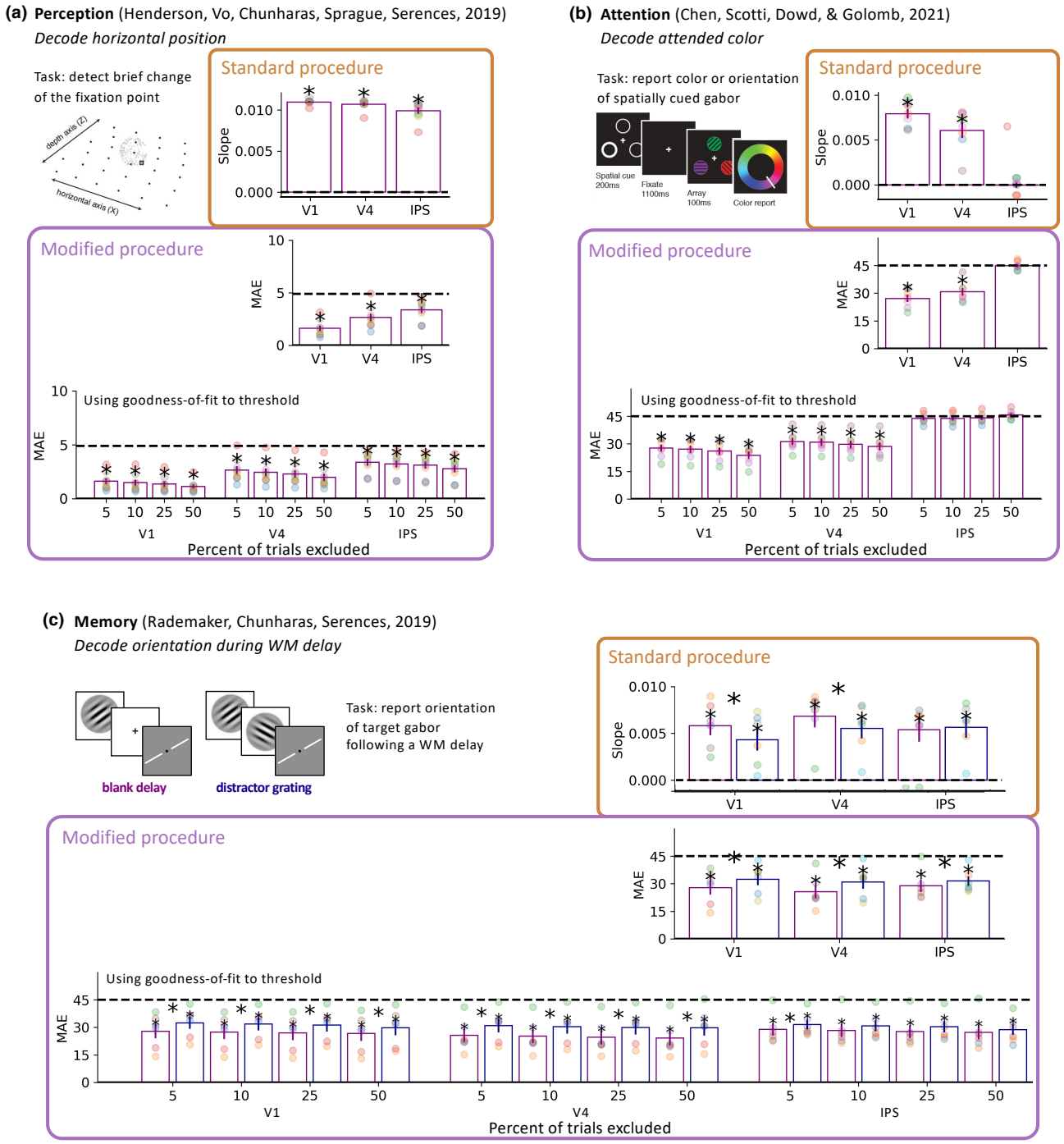

Figure 5. Results of the standard IEM procedure and our modified IEM procedure across three real fMRI datasets spanning the topics of perception (a), attention (b), and memory (c). For each dataset, results are plotted obtained from the standard procedure (orange boxes), modified procedure with full data (purple boxes, top plots), and modified procedure using increasingly stringent cutoffs based on goodness-of-fit (purple boxes, bottom plots). Bar plots depict the average slope (Standard procedure; higher is better) and MAE (Modified procedure; lower is better) across subjects, with individual subjects overlaid as colored dots. Error bars depict SEM,

dotted black lines represent chance decoding, and asterisks represent statistically significant decoding ($p<.05$). Overall results show that conclusions are similar between the standard and modified procedures, but MAE is more interpretable (not based in arbitrary units) and not prone to methodological concerns discussed in the Introduction. In addition, each dataset showed that MAE consistently improved with increasing exclusion thresholds, demonstrating the flexibility of goodness-of-fit to exclude noisy trials. See Online Methods for additional information.

Next, we tested the flexibility of the modified procedure to make use of the trial-by-trial goodness-of-fit information. For each trial, the modified procedure produces a predicted stimulus value, associated prediction error, and a goodness-of-fit value. The goodness-of-fit value is a measure of how well the predicted stimulus fits an ideal basis function centered *at that predicted value*. That is, it is a measure of the confidence of that prediction, not the accuracy of the prediction, and so is obtained independently of prediction error. To test the impact of using goodness-of-fit information on decoding performance, we performed an analysis where we excluded trials with the lowest 5%, 10%, 25%, and 50% of goodness-of-fit values (Figure 5). This resulted in visible improvements in MAE for increasing numbers of trials excluded for all three datasets (linear regression revealed significant negative slope across averaged MAEs of 0%, 5%, 10%, 25%, and 50% thresholds in all cases except for IPS in the Attention dataset). Most notably, in the Attention dataset, MAE improved with increasing confidence thresholds in V1 and V4 (where decoding was significant in the unthresholded analysis) but not in IPS (where decoding was at chance in the unthresholded analysis). Thus, the goodness-of-fit information can be used to improve decoding performance when a brain region contains reliable information about a stimulus feature, but does not produce false positives in the absence of observable stimulus-specific brain activity. These findings suggest that not only does the modified procedure produced more interpretable and

less potentially flawed results by using MAE, but statistical power can be further increased using the modified procedure by excluding trials with lower goodness-of-fit values.

3. Discussion

Inverted encoding modeling has become a popular method for predicting stimuli and investigating neural representations because of its robust performance, simplicity of linear modeling, ability to predict untrained classes, and grounding in single-unit physiology. There are various approaches researchers have employed to evaluate reconstructions, typically by averaging across trial-by-trial reconstructions and evaluating the result using metrics such as slope or amplitude. We discuss how our modifications improve the flexibility and interpretability of inverted encoding modeling while fixing important methodological concerns surrounding the standard procedure, namely how the standard procedure ignores trial-by-trial variability, does not account for the fact that a perfect reconstruction returns the basis channel, and cannot leverage uncertainty in its evaluations. The practical advantages of our method are made tangible by comparing the results of the standard procedure and our modified procedure across three real fMRI datasets, highlighting the wide range of applications intended for this modified procedure.

Importantly, our method can increase statistical power of inverted encoding modeling by leveraging uncertainty in model fits. Researchers have the flexibility to exclude trials with noisier reconstructions as assessed by evaluating how similar in shape each reconstruction is to the perfect possible reconstruction (i.e., the basis

channel) at the predicted stimulus. Note that we do not prescribe a specific cutoff for determining confidence thresholds in this paper, rather, we simply offer that such an approach is possible for increasing statistical power. For example, a researcher could weight trials with higher confidences more heavily or simply decide to exclude the noisiest 20% of trials.

Our method also improves model interpretability by evaluating reconstructions in terms of prediction error. For example, “V1 showed 10° average prediction error and V4 showed 20° average prediction error” is more interpretable than “V1 showed .02 amplitude and V4 showed .01 amplitude” because the latter is in arbitrary units, whereas MAE is in meaningful units. Further, unlike amplitude or slope, the magnitude of prediction error is not dependent on the choice of basis set and can be directly compared to other experiments using the same stimulus space.

We demonstrated the above two advantages using three real fMRI datasets. Our validations across real fMRI datasets further demonstrated how our IEM approach can be applied to both circular and non-circular stimulus spaces, is sensitive to variations in decoding performance across brain regions and experimental conditions, and can be used to accurately decode the contents of perception, attention, and internally held working memory. Our modifications allowed for the decoding performance of each dataset to be directly compared to each other and demonstrated how uncertainty, measured via goodness-of-fit, can indeed be leveraged to increase statistical power. Note that just because these three datasets produced consistent overall results (in terms of significance testing) across procedures does not ensure this will always be the

case—for less reliable results, the methodological pitfalls discussed in the Introduction become increasingly problematic for the standard procedure.

In this paper we have referred to IEMs as a specific kind of encoding and decoding model that involves simple linear regression with population-level tuning functions to decode experimental stimuli or conditions. There are more complex neuroimaging methods that can similarly be used to produce reconstructions via hypothesized tuning functions. For instance, Kay et al. (2008) decoded natural images from brain activity via voxel-level receptive field models that describe tuning functions across space, orientation, and spatial frequency. Naselaris et al. (2009) further produced Bayesian reconstructions of natural images via the combination of encoding models meant to estimate structural and semantic content. Van Bergen and colleagues (van Bergen et al., 2015; van Bergen & Jehee, 2018, 2021) introduced models where voxels with similar tuning account for shared noise and which produce trial-by-trial probability distributions such that trial-by-trial uncertainty can be obtained similarly to our procedure (although the researchers discuss this in terms of testing Bayesian theories of neural computation rather than trial thresholding). An advantage of our modified IEM procedure is that improvements to the standard IEM approach are accomplished without sacrificing simplicity—the encoding model weights and the decoding model channel responses are simply estimated via ordinary least-squares estimation.

Inverted encoding modeling has become increasingly popular in recent years, and yet the proper method for evaluating IEMs has become increasingly uncertain. As depicted in Figure 2, researchers often report IEM performance according to several

metrics because of a lack of consensus regarding the “correct” way to evaluate reconstructions. Other decoding techniques in neuroimaging, such as support vector machines or neural networks, use the easily interpretable metric of classification performance (% correct), but IEMs are typically evaluated in terms of arbitrary units that are abstracted away from the stimulus space they were intended to predict. We demonstrate a clear and practical advantage for evaluating reconstructions according to our method: researchers can increase their statistical power via thresholding, compare decoding performance across experiments, evaluate performance in stimulus space, and obtain concrete stimulus predictions (with corresponding goodness-of-fits) for every trial rather than rely on a summary statistic based in arbitrary units. Future work involving IEMs can easily adopt our modified procedure, which can be implemented via one line of code with our Python package (https://pypi.org/project/inverted-encoding; see Online Methods).

Funding

This work was funded by the National Institutes of Health (R01-EY025648 to JDG) and the National Science Foundation (NSF DGE-1343012 to PSS, NSF BCS-1848939 to JDG).

References

1. Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, 29(44), 13992-14003.
2. Brouwer, G. J., & Heeger, D. J. (2011). Cross-orientation suppression in human visual cortex. *Journal of neurophysiology*, 106(5), 2108-2119.
3. Cai, Y., Sheldon, A. D., Yu, Q., & Postle, B. R. (2019). Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. *Journal of neurophysiology*, 121(4), 1222-1231.
4. Casey, M., Thompson, J., Kang, O., Raizada, R., & Wheatley, T. (2012). Population codes representing musical timbre for high-level fMRI categorization of music genres. In *Machine Learning and Interpretation in Neuroimaging* (pp. 34-41). Springer, Berlin, Heidelberg.
5. Chen, N., Bi, T., Zhou, T., Li, S., Liu, Z., & Fang, F. (2015). Sharpened cortical tuning and enhanced cortico-cortical communication contribute to the long-term neural mechanisms of visual motion perceptual learning. *Neuroimage*, 115, 17-29.
6. Chen, J., Scotti, P. S., Dowd, E. W., & Golomb, J. D. (2021). Neural representations of task-relevant and task-irrelevant features of attended objects. *bioRxiv*. https://doi.org/10.1101/2021.05.21.445168
7. Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2), 647-660.
8. Engel, S.A., Glover, G.H., & Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex,* 7, 181–192.
9. Ester, E. F., Sprague, T. C., & Serences, J. T. (2020). Categorical biases in human occipitoparietal cortex. *Journal of Neuroscience*, 40(4), 917-931.
10. Foster, J. J., Bsales, E. M., & Awh, E. (2020). Covert spatial attention speeds target individuation. *Journal of Neuroscience*, 40(13), 2717-2726.

11. Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2017). Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychological science*, 28(7), 929-941.
12. Garcia, J. O., Srinivasan, R., & Serences, J. T. (2013). Near-real-time feature-selective modulations in human cortex. *Current Biology*, 23(6), 515-522.
13. Henderson, M., Vo, V., Chunharas, C., Sprague, T., & Serences, J. (2019). Multivariate analysis of BOLD activation patterns recovers graded depth representations in human visual and parietal cortex. *eNeuro*, 6(4).
14. Ho, T., Brown, S., Van Maanen, L., Forstmann, B. U., Wagenmakers, E. J., & Serences, J. T. (2012). The optimality of sensory processing during the speed–accuracy tradeoff. *Journal of Neuroscience*, 32(23), 7992-8003.
15. Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352-355.
16. Kim, I., Hong, S. W., Shevell, S. K., & Shim, W. M. (2020). Neural representations of perceptual color experience in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 117(23), 13145-13150.
17. Kok, P., Brouwer, G. J., van Gerven, M. A., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience*, 33(41), 16275-16284.
18. Kok, P., Mostert, P., & De Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, 114(39), 10473-10478.
19. Kok, P., Rait, L. I., & Turk-Browne, N. B. (2020). Content-based dissociation of hippocampal involvement in prediction. *Journal of Cognitive Neuroscience*, 32(3), 527-545.
20. Kok, P., & Turk-Browne, N. B. (2018). Associative prediction of visual shape in the hippocampus. *Journal of Neuroscience*, 38(31), 6888-6899.
21. Lee, H., & Kuhl, B. A. (2016). Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *Journal of Neuroscience*, 36(22), 6069-6082.
22. Liu, T., Cable, D., & Gardner, J. L. (2018). Inverted encoding models of human population response conflate noise and neural tuning width. *Journal of Neuroscience*, 38(2), 398-408.
23. Li, H. H., Sprague, T. C., Yoo, A., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *bioRxiv*. https://doi.org/10.1101/2021.04.05.438511
24. Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R., & D'Esposito, M. (2018). Flexible coding of visual working memory representations during distraction. *Journal of Neuroscience*, 38(23), 5267-5276.

25. Lorenc, E. S., Vandenbroucke, A. R., Nee, D. E., de Lange, F. P., & D'Esposito, M. (2020). Dissociable neural mechanisms underlie currently-relevant, future-relevant, and discarded working memory representations. *Scientific reports*, 10(1), 1-17.
26. Mostert, P., Albers, A. M., Brinkman, L., Todorova, L., Kok, P., & de Lange, F. P. (2018). Eye movement-related confounds in neural decoding of visual working memory representations. *eNeuro*, 5(4).
27. Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, 59(3), 2636-2643.
28. Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6), 902-915.
29. Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2), 400-410.
30. Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in cognitive sciences*, 19(10), 551-554.
31. Oh, B. I., Kim, Y. J., & Kang, M. S. (2019). Ensemble representations reveal distinct neural coding of visual working memory. *Nature communications*, 10(1), 1-12.
32. Poldrack, R. A. The physics of representation. Synthese (2020). https://doi.org/10.1007/s11229-020-02793-y
33. Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, 174, 340-351.
34. Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature neuroscience*, 22(8), 1336-1344.
35. Samaha, J., Sprague, T. C., & Postle, B. R. (2016). Decoding and reconstructing the focus of spatial attention from the topography of alpha-band oscillations. *Journal of cognitive neuroscience*, 28(8), 1090-1097.
36. Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., Tootell, R.B., 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268, 889–893.
37. Sprague, T. C., Boynton, G. M., & Serences, J. T. (2019). The importance of considering model choices when interpreting results in computational neuroimaging. *eNeuro*, 6(6).
38. Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, 24(18), 2174-2180.

39. Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, 91(3), 694-707.
40. Sprague, T. C., Itthipuripat, S., Vo, V. A., & Serences, J. T. (2018). Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. *Journal of neurophysiology*, 119(6), 2153-2165.
41. Sutterer, D. W., Foster, J. J., Adam, K. C., Vogel, E. K., & Awh, E. (2019). Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLoS biology*, 17(4), e3000239.
42. Tang, M. F., Arabzadeh, E., & Mattingley, J. B. (2019). Forward modelling reveals dynamics of neural orientation tuning to unconscious visual stimuli during binocular rivalry. *bioRxiv*, 574905.
43. van Bergen, R. S., & Jehee, J. F. (2018). Modeling correlated noise is necessary to decode uncertainty. *Neuroimage*, 180, 78-87.
44. van Bergen, R. S., & Jehee, J. F. (2021). TAFKAP: An improved method for probabilistic decoding of cortical activity. *bioRxiv*.
45. van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature neuroscience*, 18(12), 1728-1730.
46. van Moorselaar, D., Foster, J. J., Sutterer, D. W., Theeuwes, J., Olivers, C. N., & Awh, E. (2018). Spatially selective alpha oscillations reveal moment-by-moment trade-offs between working memory and attention. *Journal of cognitive neuroscience*, 30(2), 256-266.
47. Vo, V. A., Sprague, T. C., & Serences, J. T. (2017). Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. *Journal of Neuroscience*, 37(12), 3386-3401.
48. Yu, Q., & Shim, W. M. (2019). Temporal-order-based attentional priority modulates mnemonic representations in parietal and frontal cortices. *Cerebral Cortex*, 29(7), 3182-3192.
49. Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS biology*, 18(6), e3000769.

Online Methods

We performed analyses on two publicly available published datasets (Henderson, Vo, Chunharas, Sprague, & Serences, 2019; Rademaker, Chunharas, & Serences, 2019) and one unpublished dataset from our lab (Chen, Scotti, Dowd, & Golomb, 2021). Note that we only analyzed a subset of the data from each dataset, analyzing one or two conditions across three brain regions for the sake of simplicity. The experimental paradigms and conditions / regions chosen are described more in each dataset's respective subsection below.

Inverted encoding model procedures

For all datasets, we performed a set of analyses using both the standard and modified IEM procedures, as described in the Introduction, with the exception that we used iterative shifting for both the standard and modified IEM. The basis set was composed of nine equidistant channels each modeled as $\cos\left((\theta - \mu)\frac{\pi}{180}\right)^8$. We used 10-fold cross-validation, such that each iteration trained the model on 90% of the data and tested the model on the remaining 10%, repeated such that all trials were at one point decoded as part of the testing set.

For the standard IEM procedure, we aligned and averaged the single trial reconstructions into an average reconstruction and calculated slope as a traditional decoding metric. For the modified IEM procedure, we calculated absolute prediction error for each trial via the correlation table metric and then calculated MAE. We performed these steps for each subject, ROI, and condition, and then we calculated the

average slopes and MAEs across subjects. For each condition and ROI, we assessed significance via permutation testing. Significance tests were one-sided and uncorrected, calculated by comparing the t-statistic calculated from the actual data against the permuted null distribution of t-statistics (one t-statistic per each of 5,000 permutations). For the modified IEM procedure, we also repeated this analysis pipeline using varying levels of goodness-of-fit thresholds. That is, we discarded a certain percent of trials based on the worst goodness-of-fits and then calculated MAE using the remaining trials.

Perception dataset: Henderson, Vo, Chunharas, Sprague, and Serences (2019)

Data were obtained by downloading post-processed fMRI data associated with Henderson et al (2019), publicly available on OSF (https://osf.io/j7tpf/). In this experiment, nine participants attended to a central fixation while a sphere (multicolored flickering dots positioned on the shell of a 3D sphere with radius 3.4°) was presented at varying positions along the horizontal and depth axes (depth achieved through stereoscopic MR-compatible goggles). The task was to detect a brief luminance change of the fixation point. Participants completed between 7 and 21 runs, where each run of 36 trials began with a sphere presented for 3s followed by a jittered intertrial interval (2-6s). There were also runs where participants covertly attended to the sphere, but we did not include these runs in the analysis. We only reconstructed horizontal position for simplicity and because position-in-depth was only sampled across six unique locations (varied sampling across the entire stimulus space is more appropriate for inverted encoding models) whereas horizontal position was sampled across 36 unique locations (from 0.9° to 9.8° eccentricity in both directions, collapsing across position-in-depth). We

analyzed V1, V4, and IPS regions of interest which were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli (e.g., Wandell et al., 2007) while performing a covert attention task of detecting contrast dimming on a row of the checkerboard for the rotating wedge stimulus. We applied IEMs (following the procedures outlined earlier) to the post-processed data conducted by the authors of the original paper: Single-trial activation estimates consisted of averaged z-scored BOLD signal of the 3rd and 4th TRs following stimulus presentation. For more methods information, please refer to the original paper (Henderson et al., 2019).

Attention dataset: Chen, Scotti, Dowd, & Golomb (2021)

Data were previously collected in our lab for another study (Chen et al, 2021). In this experiment, seven participants completed a visual attention task. Each trial started with a central fixation cross. After 700ms, three circle outlines were displayed at equidistant locations surrounding the fixation cross for 200ms. One outline was thicker than the others, representing the spatial cue. Participants were instructed to covertly attend to the spatial cue location while maintaining fixation at the fixation cross. After 1100ms, three colored and oriented gratings were briefly displayed for 100ms, followed by a 200ms mask and a continuous color report. Participants were instructed to report the color of the grating that appeared at the location of the spatial cue. There were also trials where participants were asked to shift attention to a different spatial location before the onset of the gratings, and entire runs where participants were asked to attend and report the orientation of the grating (instead of color), but we did not include

these in our analysis. Participants completed at least 440 trials of each condition across multiple runs and sessions. We analyzed V1, V4, and IPS regions of interest: V1 and V4 were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli (e.g., Wandell et al., 2007), while IPS was defined from the Destrieux atlas (Destrieux, Fischl, Dale & Halgren, 2010) in Freesurfer (parcel labelled "S_intrapariet_and_P_trans"). To obtain single-trial neural activations for IEM, we modified a commonly used single-trial general linear model (GLM) approach (Mumford et al., 2012) to improve the model sensitivity and account for the large number of trials. Specifically, we conducted 40 GLMs per subject, where each GLM includes one regressor per run for one of the 40 trials in that run and one regressor per run for all the other remaining trials in that run. In this way, across the 40 GLMs, each trial in the experiment had an estimated single-trial beta weight. For more methods information, please refer to the original paper (Chen et al., 2021).

Memory dataset: Rademaker, Chunharas, and Serences (2019)

Data were obtained by downloading post-processed fMRI data associated with Rademaker et al (2019), publicly available on OSF (https://osf.io/dkx6y). We analyzed Experiment 1 of Rademaker et al. (2019), where six participants underwent a visual working memory task. For each trial, a cue indicating the distractor condition was shown for 1.4s, followed by a target grating shown for .5s where participants were instructed to memorize its orientation, followed by a 1s blank delay, and then an 11s delay where 3 possible distractor conditions were possible: blank delay, Fourier-filtered noise, or distractor grating of a pseudo-random orientation. Following an additional 1s blank

delay, participants had 3s to report the orientation of the target grating, and finally a variable intertrial interval (3/5/8s). Each participant completed 108 trials per distractor condition. We only reconstructed the blank delay and distractor grating conditions for simplicity. We analyzed V1, V4, and IPS regions of interest which were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli (e.g., Wandell et al., 2007). We applied IEMs to the post-processed data conducted by the authors of the original paper: Single-trial activation estimates consisted of averaged BOLD signal between 5.6-13.6s (7-17 TRs) after target onset. For more methods information, please refer to the original published paper (Rademaker et al., 2019).

Python package: inverted-encoding

We have released the Python 3 package “inverted-encoding” on PyPi (https://pypi.org/project/inverted-encoding/) and GitHub (https://github.com/paulscotti/inverted_encoding) for easy implementation of our modified inverted encoding model procedure. The package contains two main functions, “IEM” and “permutation.”

For the “IEM” function, the only necessary inputs are an array of the stimulus features for every trial and a trial by voxel activations matrix (note: inputs other than voxels may be used for other modalities). The basis set can be specified as an optional parameter and will otherwise default to a basis set composed of nine equidistant channels each modeled as $\cos\left((\theta-\mu)\frac{\pi}{180}\right)^8$. The stimulus space defaults to a circular 0-179° range but can be optionally set to other ranges. Non-circular stimulus spaces

can be set by the Boolean parameter “is_circular.” The IEM procedure defaults to a 10-fold cross-validation procedure but can be optionally specified. The final outputs are an array of each trial’s predicted stimulus and an array of each trial’s corresponding goodness-of-fit. The user can then compute MAE themselves by averaging the (circular) absolute error between the predicted stimulus features and the actual stimulus features. The user can decide whether they want to threshold any trials using the provided goodness-of-fit values prior to calculating MAE.

For the “permutation” function, the only necessary input is an array of the actual stimulus features. For each iteration, the stimulus features are randomly shuffled and used as the predicted stimuli to compute the MAE. The function outputs a null distribution of MAE values for the user to compare against the MAE obtained from the “IEM” function. A more exact and computationally intensive method would be to rerun the entire IEM pipeline with shuffled stimulus labels on every iteration to obtain the null distribution. This can also be performed using our package by simply repeating the IEM function with a different shuffling of the stimulus features for every iteration. Our exploratory comparisons of null distributions obtained using both approaches across the three fMRI datasets discussed in the main text yielded no obvious differences.