PLOS ONE

RESEARCH ARTICLE

Inferring exemplar discriminability in brain representations

Hamed Nili[1]*, **Alexander Walther**[2], **Arjen Alink**[3], **Nikolaus Kriegeskorte**[4]

1 Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, England, United Kingdom, **2** MRC Cognition and Brain Sciences Unit, Cambridge, England, United Kingdom, **3** University Medical Center Hamburg-Eppendorf, Hamburg, Germany, **4** Zuckerman Institute, Columbia University, New York, NY, United States of America

* nili.hamed@gmail.com, hamed.nili@mrc-cbu.cam.ac.uk

Citation: Nili H, Walther A, Alink A, Kriegeskorte N (2020) Inferring exemplar discriminability in brain representations. PLoS ONE 15(6): e0232551. https://doi.org/10.1371/journal.pone.0232551

Editor: J Malo, Universitat de Valencia, SPAIN

Received: January 23, 2020

Accepted: April 16, 2020

Published: June 10, 2020

Data Availability Statement: The data that support the findings of this study are available on https://osf.io/dxpba/.

Funding: This work was supported by the UK Medical Research Council and by a European Research Council Starting Grant (261352) and Wellcome Trust Project Grant (WT091540MA) to NK, a Cambridge Overseas Trust scholarship to HN, a Gates Cambridge Scholarship to AW and a British Academy postdoctoral fellowship to AA.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Representational distinctions within categories are important in all perceptual modalities and also in cognitive and motor representations. Recent pattern-information studies of brain activity have used condition-rich designs to sample the stimulus space more densely. To test whether brain response patterns discriminate among a set of stimuli (e.g. exemplars within a category) with good sensitivity, we can pool statistical evidence over all pairwise comparisons. Here we describe a wide range of statistical tests of exemplar discriminability and assess the validity (specificity) and power (sensitivity) of each test. The tests include previously used and novel, parametric and nonparametric tests, which treat subject as a random or fixed effect, and are based on different dissimilarity measures, different test statistics, and different inference procedures. We use simulated and real data to determine which tests are valid and which are most sensitive. A popular test statistic reflecting exemplar information is the *exemplar discriminability index (EDI)*, which is defined as the average of the pattern dissimilarity estimates between different exemplars minus the average of the pattern dissimilarity estimates between repetitions of identical exemplars. The popular across-subject *t* test of the EDI (typically using correlation distance as the pattern dissimilarity measure) requires the assumption that the EDI is 0-mean normal under H_0. Although this assumption is not strictly true, our simulations suggest that the test controls the false-positives rate at the nominal level, and is thus valid, in practice. However, test statistics based on average Mahalanobis distances or average linear-discriminant *t* values (both accounting for the multivariate error covariance among responses) are substantially more powerful for both random- and fixed-effects inference. Unlike average cross-validated distances, the EDI is sensitive to differences between the distributions associated with different exemplars (e.g. greater variability for some exemplars than for others), which complicates its interpretation. We suggest preferred procedures for safely and sensitively detecting subtle pattern differences between exemplars.

1. Introduction

Brain representations are increasingly investigated with pattern-information analyses of data acquired with brain imaging or neuronal recording techniques [1–11]. Information carried by brain response-patterns can be explored at different levels. Two common levels are the category and the exemplar level. Category information has previously been studied using regional average activation (e.g. [12]) and pattern-decoding approaches (e.g. [1]). Pattern-classifier decoding lends itself naturally to investigating category information with the category labels being decoded from the brain-activity patterns. Recent studies have estimated a unique response pattern for each individual stimulus and investigated not only category information, but also within-category exemplar information [6,13,14]. Exemplar information is present to the degree that different exemplars elicit distinct representational patterns. Within-category effects are often subtle, thus powerful tests are needed to detect them. Here we use the term "exemplar discriminability" generically to denote the average discriminability across all pairs among a set of experimental conditions.

Studying representational distinctions within categories can arise at different contexts. For example, in studies of visual face representations, it is important to quantify to what extent a face region distinctly represents individual faces [15–17]. Classifier decoding can be used to test for exemplar information. For example, a classifier can be trained to distinguish two exemplars within a category. For power, we would then like to have many repetitions of each stimulus. This would suggest repeating a small number of exemplars many times in the experiment (e.g. [16]). It is also desirable, however, to sample the category with many different exemplars in order to get a richer description of its underlying representation. There is a trade-off between the number of stimuli and the number of repetitions of each stimulus, because the total time available for measuring brain activity in a subject is usually limited. If we have many different exemplars (i.e. a condition-rich design), we can typically repeat each stimulus only a few times. This severely limits our power to detect the discriminability of a given pair of exemplars. In the case of condition-rich designs, it is therefore desirable to pool the evidence across many pairs of exemplars. This is achieved by a summary statistic that combines the evidence of discriminability across all pairs of exemplars.

A popular summary statistic reflecting exemplar discriminability among a set of experimental conditions from the same category is the "exemplar discriminability index" (EDI; e.g. [18–25]). The EDI is defined as the average between-exemplar dissimilarity estimate minus the average within-exemplar dissimilarity estimate (Fig 1). A within-exemplar dissimilarity is a dissimilarity between independent measurements of the activity pattern elicited by repeated presentations of the same stimulus. The average within-exemplar dissimilarity estimate thus reflects the noise in the measurements. Subtracting it is essential when the dissimilarity estimate (i.e. pattern-distance estimates) is positively biased. The correlation distance, for example, which was used in most of the cited studies, is non-negative by definition, and therefore positively biased (although correlation coefficients are between -1 and 1, correlation distance, defined as 1 minus the correlation coefficient, is between 0 and 2). Subtracting the average within-exemplar dissimilarity removes the bias and enables inference. Typically, the EDI is computed for each subject and tested at the group level using a one-sided *t* test, treating subject as a random effect. It must be noted that the previous studies that used EDI do not use this nomenclature. However, we consistently use 'EDI' for any summary statistic computed in the way that is described above, i.e. any measure of exemplar discriminability.

The cited studies using this approach rely on splitting the data into two independent halves (*e.g.* odd and even runs in fMRI measurements) and comparing response patterns between the two halves. The between-halves dissimilarities are assembled in a split-data representational

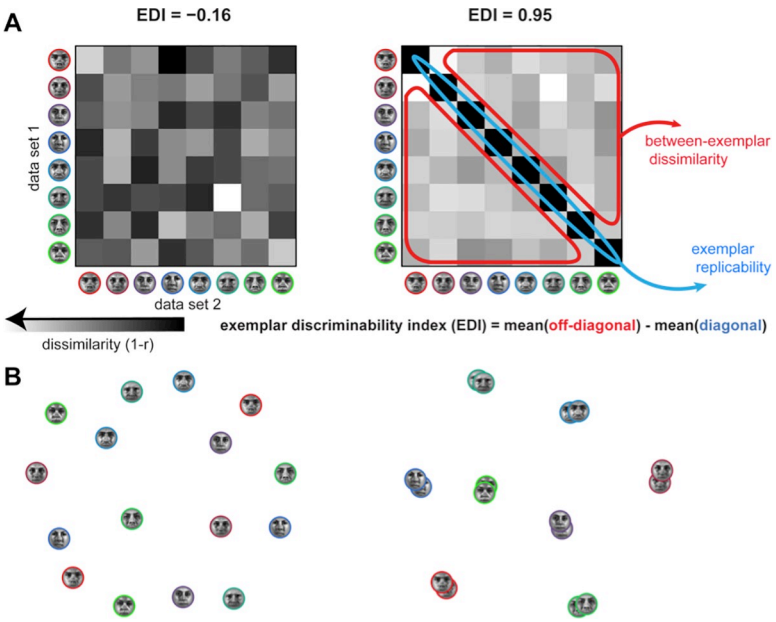

Fig 1.

https://doi.org/10.1371/journal.pone.0232551.g001

dissimilarity matrix (*sdRDM*), which is indexed vertically by response patterns estimated from data set 1 and horizontally by response patterns estimated from data set 2. Each entry of the sdRDM contains one dissimilarity between two response-pattern estimates spanning the two data sets. The order of the exemplars is the same horizontally and vertically, so the diagonal of the sdRDM contains the within-exemplar pattern dissimilarity estimates (Fig 1A).

The within-data-set pattern dissimilarities are not used for either within- or between-exemplar pattern comparisons. This is important because patterns measured closer in time tend to be more similar due to measurement artefacts [26, 27].

Note that for the popular correlation distance (1—Pearson r), the difference of distances is equal to the negative of the difference of correlations: d1 –d2 = (1-r1)—(1-r2) = r2—r1. For a consistent comparison with other measures of pattern dissimilarity, however, we use the correlation *distance* here.

The EDI *t* test has some caveats. First, this approach only allows testing exemplar information with subject as a random effect. Single-subject or group-level inference with subject as fixed effect cannot be accommodated in this approach. Second, it is possible that the assumptions of the test are not met. A *t* test requires that the data are normally distributed under H_0. The EDI is the difference of two average dissimilarities. For non-negative dissimilarities like the Euclidean distance, the distributions of the within- and the between-exemplar dissimilarities are skewed (limited by 0 on the left, unlimited on the right). Under the null hypothesis, the within- and between-exemplar dissimilarities are all samples from the same distribution (i.e. for a given exemplar, the effect of changing the exemplar is not different from the effect of having another measurement for that exemplar). The expected value of the mean of the diagonal (within-exemplar) entries of the sdRDM is therefore equal to the expected value of the mean of the off-diagonal (between-exemplar) entries. The expected value of the difference between those means, i.e. the expectation of the EDI, is therefore also zero under H_0. However, there are more between-exemplar than within-exemplar dissimilarities. For N exemplars, the sdRDM has N^2 entries, so there are N diagonal entries (within-exemplars) and $N^*(N\text{-}1)$ off-diagonal entries (between-exemplars). Under H_0, the distribution of the average between-

exemplar dissimilarities will therefore be narrower than that of the average within-exemplar dissimilarities. The EDI, thus, is a difference between two random variables, which have different variances and skewed distributions. The null distribution therefore need not be symmetric and can be non-Gaussian. The *t* test, therefore, is technically invalid as a test of the EDI.

Fig 2 illustrates different scenarios for the distributions of the diagonal and off-diagonal means under H_0. The distributions can be symmetric or skewed; furthermore, they could have the same shape (implying also the same width) or different shapes (e.g. different widths). Note that a difference between two random variables is symmetrically distributed about 0 if either (a) the variables are identically distributed (can be skewed in this case) or (b) each of the variables is symmetrically distributed about the same expected value (they can have different shapes, e.g. different variances, in this case). However, the diagonal and off-diagonal means are both skewed and non-identical (different variances), so the EDI is not symmetrically distributed under H_0 thus not normally distributed. Therefore, the assumptions of the *t* test are not strictly met.

This paper has three aims. First, we assess the practical validity of the popular EDI *t* test by simulation. Second, we introduce a number of alternative tests that are valid and, unlike the EDI *t* test, enable single-subject inference and group-level fixed-effect inference, along with random-effects inference. Third, we compare all these tests in terms of their power to detect exemplar discriminability in real data from functional magnetic resonance imaging (fMRI). The simulations show that the theoretical violations of the assumptions of the EDI *t* test are minimal in practice and hence the test is generally valid. However, exemplar discriminability tests based on dissimilarity measures that account for multivariate error covariance (Mahalanobis distance and linear discriminant *t* values; [4, 16, 28, 29]) are substantially more powerful and enable inference for single subjects.

2. Material and methods

2.1. fMRI experiment

17 participants (10 female, age range 20–38) underwent four functional runs of scanning in two separate scanning sessions. In each run, participants were presented with 24 images of real-world objects belonging to two categories with 12 objects in each. Categories were changed from run to run towards ever more fine-grained categorical distinctions: animate/inanimate, faces/bodies, animal faces/human faces, and male faces/female faces. Participants were instructed to either categorize a stimulus based on the one previously shown (session one) or to complete a visual fixation task (session two). Each session also included two runs during which participants viewed retinotopic mapping stimuli (for details see [26]) and images depicting faces, houses, objects and scrambled objects. The analysis of brain responses to these stimuli allowed us to define regions of interest for FFA, PPA, LOC and early visual areas V1-3. Functional EPI images covering the entire brain were acquired on a 3T Siemens Trio scanner using a 32-channel head coil (32 slices, resolution = 3mm isotropic, inter slice gap = 0.75mm, TR = 2000ms). For each participant we also obtained a high-resolution (1mm isotropic) T1-weighted anatomical image using a Siemens MPRAGE sequence.

All participants gave their informed consent after being introduced to the experimental procedure in accordance with the Declaration of Helsinki. The experimental procedure has been approved by the Cambridge Psychology Research Ethics Committee (ethics reference number: CPREC 2010.52)

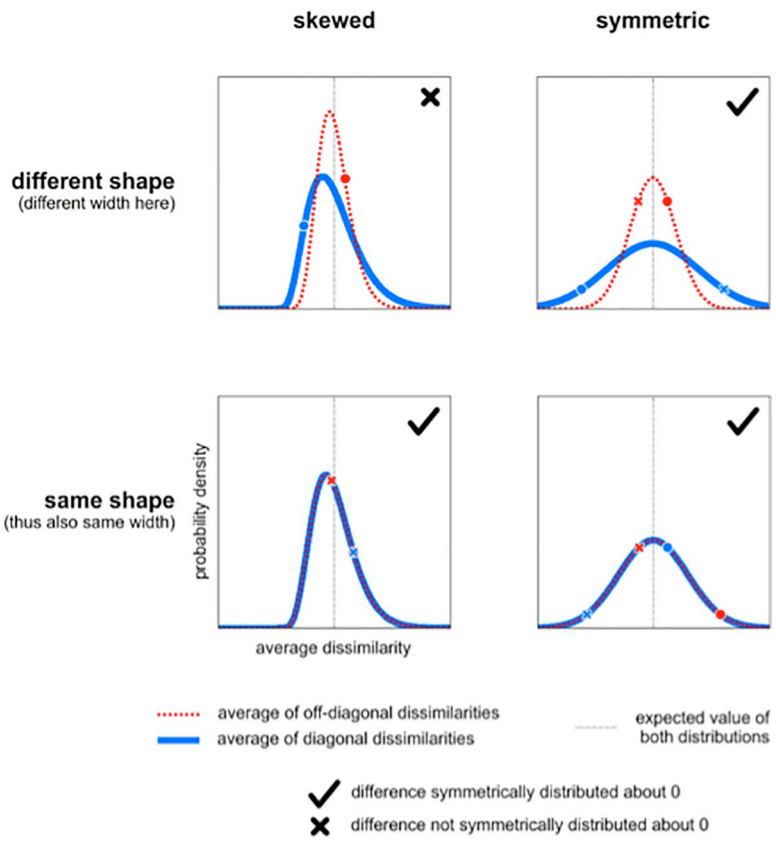

Fig 2.

https://doi.org/10.1371/journal.pone.0232551.g002

2.2. Tests statistics for measuring exemplar discriminability

2.2.1. Exemplar Discriminability Index (EDI). Representational similarity analysis (RSA, [30]), characterizes the representations of a brain region by a representational dissimilarity matrix (RDM). An RDM is a distance matrix composed of distances between response patterns corresponding to all pairs of experimental conditions. In cases where the responses to the same exemplars are measured in two independent data sets, a similar approach can be taken and the response patterns can be compared in a split-data RDM (sdRDM, Fig 1A). Rows and columns of an sdRDM are indexed by the exemplars in the same order. Entries of this type of RDM are comparisons between responses to the same or different exemplars in two different data-sets.

The EDI is the difference of the average between-exemplar and the average within-exemplar dissimilarities. Pattern dissimilarity (e.g. elements of the sdRDM) can be measured using various distance measures. In this paper, we explore the following measures of pattern dissimilarity:

- *Euclidean distance*: For two vectors **a** and **b,** the Euclidean distance is the L^2 norm of the difference vector **a**-**b**. Geometrically, it corresponds to the length of the vector that connects the two vectors **(a** and **b**) to each other.
- *Pearson correlation distance*: The Pearson correlation distance for two vectors **a** and **b** is equal to 1 minus the Pearson correlation coefficient between them. The correlation distance

has a geometrical interpretation: It is one minus the cosine of the angle between the mean-centered vectors of **a** and **b** (e.g. voxel-mean centered transformations of **a** and **b**).

- *Mahalanobis distance*: The Mahalanobis distance is the Euclidean distance between the two vectors after multivariate noise normalisation. Multivariate noise normalisation is a transformation that renders the noise covariance matrix between the response channels identity. In this transformation, the response patterns are normalised through scaling with the inverse square root of the error-covariance. Therefore, if we have a response matrix, **B** (Response matrix is a matrix of the distributed responses to all exemplars, with each row containing the response to one exemplar in all response channels. The size of this matrix will be number of exemplars by number of response channels), we can noise-normalise it like so:

$$\mathbf{B}^{*} = \mathbf{B}\hat{\Sigma}^{-\frac{1}{2}} \quad \text{Eq1}$$

where **B** is the original response matrix, $\hat{\Sigma}$ is the estimated error variance-covariance matrix (square matrix with number of rows and columns equal to the number of response channels, e.g. voxels), and $\mathbf{B}^{*}$ is the response matrix formed from the response patterns after multivariate noise normalisation (number of exemplars by number of response channels). If the number of response channels greatly outnumbers the number of exemplars, $\hat{\Sigma}$ will be rank-deficient and hence non-invertible. To ensure invertibility, we regularize the sample covariance estimate with optimal shrinkage [31]. This method shrinks the error covariance matrix towards the (invertible) diagonal matrix using an optimally weighted linear combination of the sample covariance matrix and the diagonal covariance matrix estimate. The optimal shrinkage minimizes the expected quadratic loss of the resultant covariance estimate.

2.2.2. Average LD-*t* / LDC. Linear discriminant analysis (LDA) could also be used to estimate the discriminability values between pairs of exemplars. Similar to EDI, this method also relies on having two independent measurements (i.e. two datasets) for the same set of exemplars. For each exemplar pair, the Fisher linear discriminant is fitted based on the data from one of the splits. The data from the other split is then projected onto the discriminant line. The linear discriminant *t* value (LD-*t*, [16, 28, 29]) would then be obtained by computing the *t* value for data from that pair after projection onto the discriminant line. Under the null hypothesis, the LD-*t* is symmetrically *t*-distributed around zero. Therefore, it is possible to make inference on mean discriminabilities across many pairs of stimuli by performing either fixed effects analysis or across-subjects random-effects tests. The LD-*t* can be interpreted as a cross-validated and noise-normalized Mahalanobis distance [28, 29]. The average *t* value of all exemplar pairs would then be a measure of the discriminability of all exemplars.

The discriminabilities for any two pairs of exemplars are not independent (e.g. discriminabilities of the same exemplar with two different ones). The standard error of the average *t* value is therefore smaller than 1 (the standard error of a standard *t* value), but larger than $1/\sqrt{n}$ where n is the number of all averaged *t* values (one for each pair). Therefore, we understand that treating the average LD-*t* as a proper t value would be conservative (In fact the LD-*t* values are already averaged across the folds of crossvalidation, making such a test even more conservative.).

One could alternatively use the average linear discriminant contrast (LDC), which is also known as the *crossnobis* estimator (squared cross-validated Mahalanobis distance), as a measure of exemplar disciminability [28, 29].

2.2.3. EDI or average LD-*t* after removing the effect of univariate activation from the dissimilarity measures. Response-pattern dissimilarities could be influenced by differences

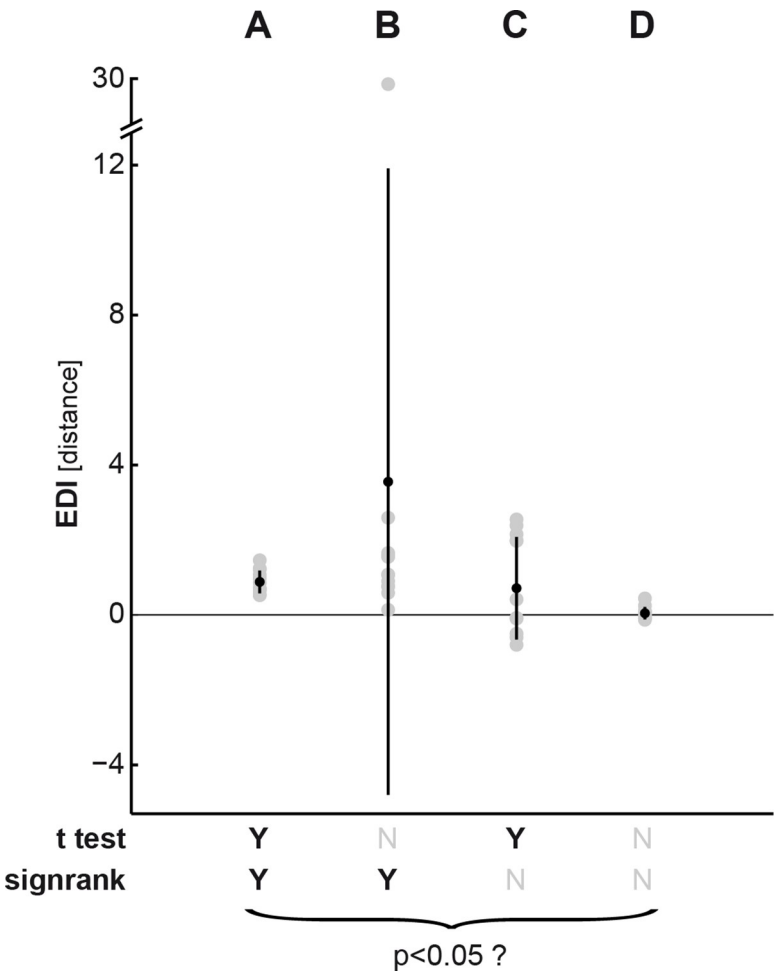

Fig 3.

https://doi.org/10.1371/journal.pone.0232551.g003

in the average activation of a brain region. Removing the contribution of the activation differences might be interesting in scenarios where only pattern differences are investigated.

The correlation distance for any pair of conditions is 1 minus the inner product of the standardized activity patterns after subtracting the regional average activation from each. To remove the effect of univariate activation from the Euclidean and Mahalanobis distance, we computed the distance measure after subtracting the across-response-channels mean from the pattern of each condition (results from these are denoted as "average activation removed" in Figs 8 and 10)

$$\tilde{\mathbf{b}}_i = \mathbf{b}_i - \bar{b}_i \mathbf{1} \qquad \text{Eq2}$$

where $\underline{b}$i is the average value of all response channels for condition i and 1 is a 1 by number of response channels row vector containing only ones. The LD-t is computed as explained in Nili et al. [28]. However, after estimating the discriminant weights from the training data, we subtract the component that is along 1 from the weight vector. The reasoning is that the component along 1 corresponds to the overall differences in activation for the two conditions, hence removing it from the weights will result in a LD-t value with no contribution from activation

differences. More specifically, we estimate the weight vector, w, from the training data (e.g., dataset 1) and replace it with $\mathbf{w}-\langle\mathbf{w},\mathbf{1}\rangle$, where $\langle\rangle$ denotes the inner product.

2.3. Tests for inference

2.3.1. Subject as random effect. Treating subjects as a random effect allows inference about the population from which the subjects were randomly drawn. For random-effects analysis of EDI values, the summary statistic is first estimated in individual subjects. Estimates of all subjects are then jointly tested using either a one-tailed *t* test or a one-tailed Wilcoxon signed-rank test.

2.3.1.1. One-sided t test. The standard method for testing EDIs is the *t* test. Since there is a hypothesis about the direction of the EDI (i.e. positive EDI for exemplar information), a one-tailed test is appropriate. The *t* test assumes that the data come from a population that is normally distributed under the null hypothesis (H_0).

2.3.1.2. One-sided Wilcoxon signed-rank test. In cases where the assumption of normality of the data seems unreasonable, one might consider using non-parametric alternatives to the *t* test. In this paper, we consider the Wilcoxon signed-rank test [32] as the non-parametric alternative. This non-parametric statistical test compares the median of its input to zero. In this test the EDIs are first ranked according to their absolute values. The difference of the ranks for the positive and negative EDIs is then computed. The p-value corresponding to the difference of the ranks is the output of the test.

Another valid nonparametric test is the sign-test [33]. Significance of the sign-test is merely based on the number of positive and negative values in a given sample. (given the number of positive and negative samples, the p-value would be directly computed from the Bernouli distribution). Therefore, the sign-test discards any information about the magnitude of the samples and is not considered an appropriate non-parametric alternative for *t test* in testing EDIs.

2.3.1.3. The t test and Wilcoxon signed-rank test are affected by different properties of the EDI distribution. *t* test and Wilcoxon signed-rank test base their inference on different statistics of the data distribution: while the former computes the mean and infers the standard error around it based on the sample variance, the latter computes a tail probability that relates to the sum of positive ranks of the data. These characteristics are independent from one another, therefore the tests may yield considerably different p-values depending on the degree to which each feature is pronounced in the data. To illustrate this, we simulated four sets of 12 EDIs, each by drawing random data points from four Gaussian distributions with different means and variances (Fig 3). We then submitted each set to a *t* test and a Wilcoxon signed-ranked test (both right-tailed and testing against the null hypothesis that the data come from a zero-mean distribution) and thresholded the resulting p-values by the conventional $p<0.05$ criterion.

In set A, the mean of the sample is above zero and the sample variance is small; therefore, both tests yield a significant p-value. In set B, the sample mean is more positive than in set A. However, although all EDIs are positive, an outlier in the sample (value close to 30) drastically inflates the sample variance. This leads to a non-significant p-value when applying the *t* test, because the outlier increases the standard error and therefore diminishes the *t* value. By contrast, the Wilcoxon signed-rank test is not drastically affected by the outlying EDI and yields a significant p-value. The reason for this is that while the outlier has the highest rank, the rank does not contain any information on how far away this value is from the sample mean. The signed-rank test is therefore more robust against extreme outliers than the *t* test.

In set C, the variance is larger than in set A, but deviations from the sample mean are still relatively small. Therefore, the *t* test returns a significant p-value, indicating the mean EDI of

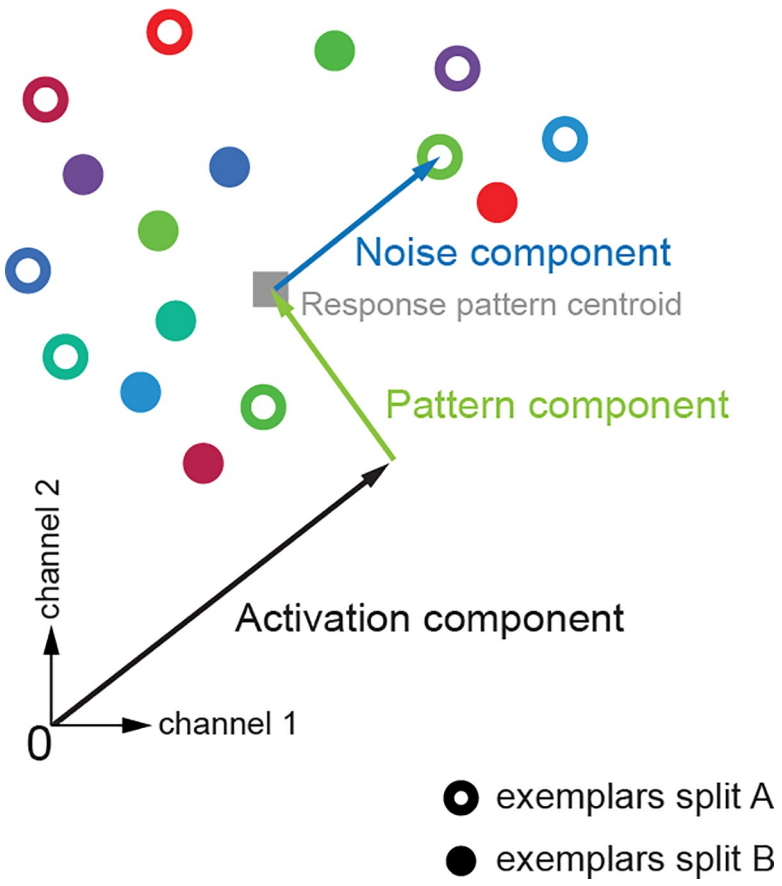

Fig 4.

https://doi.org/10.1371/journal.pone.0232551.g004

the sample is different from zero, suggesting exemplar information. On the contrary, submitting the same set of data points to the Wilcoxon signed-rank test yields a p-value that is above the significance threshold. The reason for this is that a substantial proportion of EDIs in the sample are negative, making the tail probability too small to result in a significant p-value. In this case the *t* test is the more sensitive test because it accounts for the overall sample variance rather than ranks. Finally, in set D the mean is close to zero, hence both tests do not reject the null hypothesis.

2.3.2. Single-subject or subject as fixed effect. In clinical cases or basic research, it is sometimes desirable to perform tests at the single-subject or group-average level. For example, we may want to test if familiar faces are represented distinctly in one subject or in a particular group (e.g. patients). These tests do not require generalization to the larger population from which the current subjects are sampled from. We suggest two tests for the fixed effect analysis of exemplar information.

2.3.2.1. RDM-level condition-label randomisation test for EDI. Under the null hypothesis, the within-exemplar and between-exemplar dissimilarities are exchangeable. Therefore, for a given split-data RDM (single-subject sdRDM or group-average sdRDM, for single-subject or group-level fixed-effect analysis, respectively), one can independently permute the rows or the columns of the sdRDM many times and compute the EDI at each iteration. The p-value for the non-parametric test is then the proportion of more extreme (greater) values in the null distribution compared to the EDI for the single subject or group average. So if N denotes the null

distribution of EDIs and EDI_m the single-subject or group-average EDI, the p-value is obtained according to the following formula:

$$p = \frac{n(N > EDI_m)}{n(N)} \quad \text{Eq3}$$

where *n()* is an operator that counts the number of elements in a set (i.e. its cardinality).

This test can be efficiently implemented by using the mean of the diagonal entries as the test statistic. Since the sum of the diagonal and off-diagonal entries is constant across permutations, one can equivalently compute the p-value as the proportion of diagonal averages (across many permutations) that are smaller than the diagonal average for the single subject or group.

2.3.2.2. Pattern-level condition-label randomization test for average LD-t. In contrast to sdRDMs, the within-exemplar distances are not estimated in LD-*t* RDMs and the exchangeability (under H_0) could not be applied in the same way. However, one can still estimate the null distribution of the EDI by computing the LD-*t* RDMs under the null hypothesis for one subject or group of subjects. For single-subject inference, we fit the discriminant line for the training dataset (*e.g.* dataset 1) and test it on the remaining data (*e.g.* dataset 2) after condition-label randomisation (randomly shuffled test-data). We then obtain the null distribution of the average LD-*t* values by aggregating the results from the null LD-*t* RDMs (*i.e.* LD-*t* RDMs obtained under the null hypothesis for different iterations). Fixed-effect group-level analysis is carried out in the same way as single-subject analysis. The null LD-*t* values are estimated for each subject and the null distribution for group-level analysis will be the distribution of subject-averaged values. Note that this test requires more computations than the RDM-level condition label randomisation test.

2.4. Scenarios for assessing the statistical tests

A statistical test can lead to errors in two cases: 1) When the data comes from the null distribution and the test gives significant results (false positives, *type I error*). 2) when the alternative hypothesis holds but the test does not detect it *(false negatives*, *type II error)*. If the type I error is large, the test lacks *specificity* and is not valid. If the type II error is large, the test lacks *sensitivity* and is not powerful. Ideally one would seek a test that is both sensitive and specific. In this section, we first test the validity of the underlying assumptions for the *t* test of EDIs and then investigate the specificity and sensitivity of different EDI tests.

2.4.1. H0: Simulation. Simulations allow us to simulate multivariate ensemble vectors with known properties. We use a number of parameters to simulate multi-normal activity patterns under H0 for two datasets in each simulated subject. For any point in the parameter space (*i.e.* any combination of parameters), we simulated patterns for a large number of subjects (10,000 subjects) for each dataset. The distribution of the EDIs is then a good estimate of the EDI null distribution for that particular set of parameter values. The main purpose of this rich estimation of the EDI null distribution is to assess the validity of the required assumptions for the *t* test. For the conventional *t* test approach to give interpretable results, the null distribution of the population needs to be zero-centered and reasonably Gaussian.

Fig 4 illustrates our simulation setup. For any combination of parameters, we apply three different tests to the simulated null EDIs:

- *One-sided Wilcoxon signed-rank test*: Tests if the median of the EDI null distribution is different from zero
- *Lilliefors test of Gaussianity*: Tests if the EDI null distribution is Gaussian. The Lilliefors test [34] is a normality test with its H_0 being that the data is normal (unknown mean and

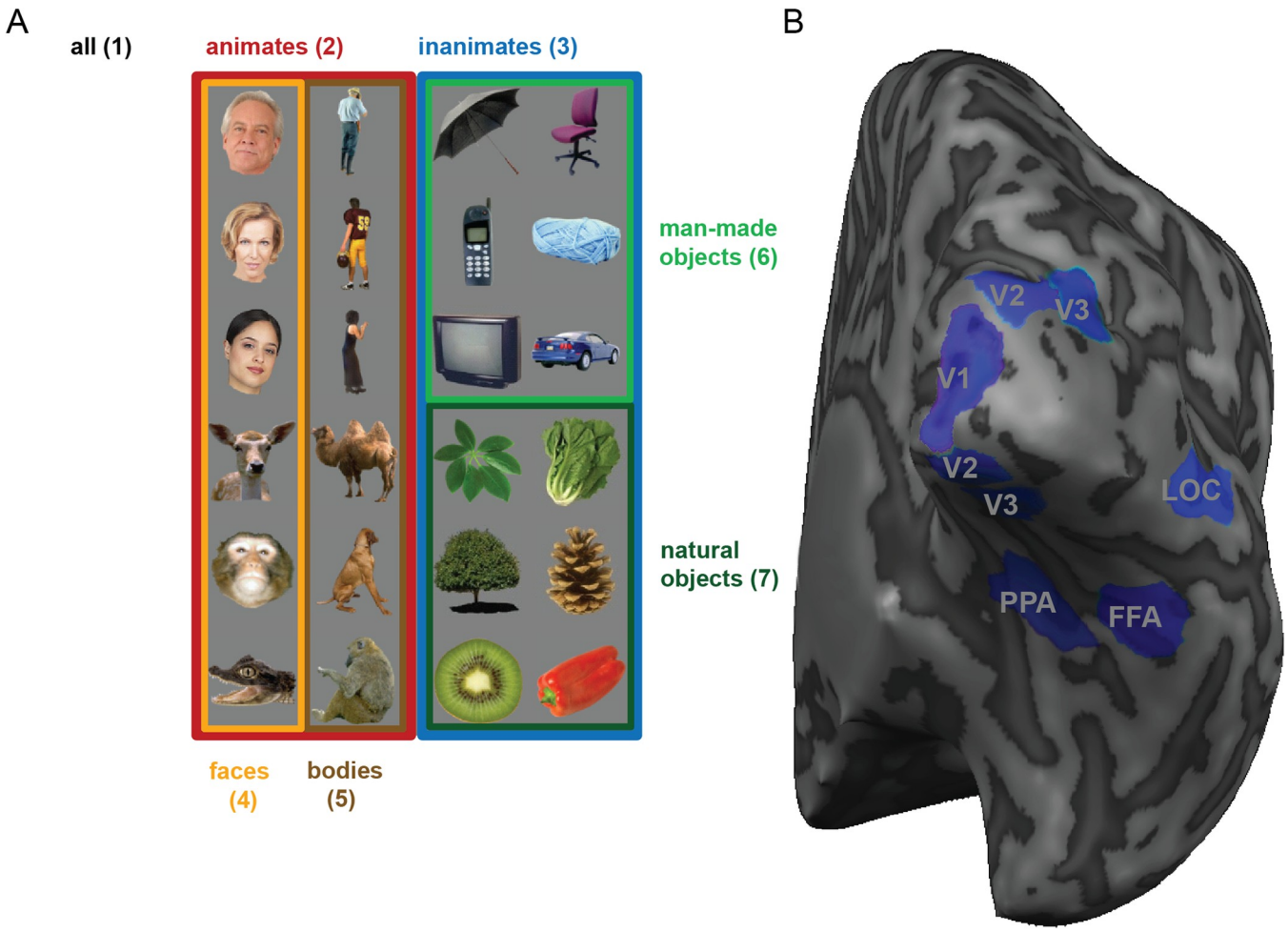

Fig 5.

https://doi.org/10.1371/journal.pone.0232551.g005

standard deviation). This test is based on the maximum discrepancy between the empirical distribution function and the cumulative distribution function of the normal distribution with the estimated mean and estimated variance.

- *One-sided t test*: Tests if the mean of the EDI null distribution is different from zero. Considering this allows us to estimate the false-positive rate of the standard *t* test approach for every point of the parameter space.

As illustrated in Fig 4, the simulated multivariate responses were based on five parameters. All simulated patterns can be imagined as vectors in a space with as many dimensions as response channels (e.g. voxels in fMRI). The *activation component* is a univariate component (emanating from the origin) that affects all response channels of all exemplars equally in both datasets. The *pattern component* is then the variance of activity in each response channel that is added to the activation component. Here, zero pattern component variance means that the centroid is on the all-ones vector, i.e. the response is equal in all response channels and equal to the grand mean, hence there is no spatial variability across the simulated response channels for the centroid. Conversely, a high variance means that the average pattern across splits and

exemplars has great spatial variability. Adding activation and pattern component gave the response pattern centroid. Response patterns for two data splits of individual exemplars were then simulated by taking random samples from a Gaussian distribution centred at the centroid with a certain noise variance for each data split (denoted by the *noise component*). This noise variance parameter therefore determined the variability of the exemplars within and between repetitions. This is sensible because under the null hypothesis, all exemplars are indiscriminable and the within- and between-exemplar variabilities are equal.

The other two parameters are the *number of response channels* and the *number of exemplars*. Consider an fMRI experiment investigating whether the representations of a brain region distinguish between different face images. In that case, the number of exemplars would correspond to the number of face images that were presented to each participant and number of response channels would be the number of voxels whose activities were recorded during the scan and were investigated. Note that in practice the values of these two parameters are chosen by the analyst/experimenter. One can intuitively think that the number of exemplars may play an important role. While having more exemplars reduces the sampling error by having richer samples of the stimulus space, it also increases the gap between the number of diagonal and off-diagonal estimates of an sdRDM. Therefore, one can speculate that more exemplars may not be as advantageous for the *t* test, because this also makes it more likely that the underlying assumptions of the test are violated. Exploring the effect of the number of response channels is also important. It would be essential to know how exemplar information in distributed patterns depends on the number of response channels. For example, in fMRI analysis, researchers often replicate the same effect for a range of ROI sizes (e.g., [6]). This analysis helps reveal potential special *peculiar* effects for some number (or range of numbers) of response channels that are due to sensitivity of the tests or validity of the required assumptions and not the properties of the distributed representations.

Exploring the parameter space for multiple simulation settings can lead to false positives due to the testing of multiple hypotheses (each test is repeatedly applied for all possible combinations of the parameters). To control the type-I error rate we keep the false discovery rate (FDR) at 5% [35]. Moreover, we also report results from applying the more conservative Bonferroni correction for multiple comparisons (i.e., controlling the family-wise error rate).

2.4.2. H0: Simulation by shuffling fMRI data. The null hypothesis, H_0, can also be simulated from real fMRI data. We simulate H_0 from data at the single-subject level and obtain group data under H_0 by aggregating the simulated data from all subjects. For tests that rely on statistics obtained from an sdRDM (e.g. EDI based on Pearson correlation distance), null-sdRDMs are obtained in each subject by permuting the rows and columns independently (note that exchangeability holds for H_0). For tests that rely on the LD-*t* values, the exchangeability is applied by randomly permuting the order of exemplar predictors in the design matrices of both data splits. The group aggregate is estimated for a large number of H_0 iterations and each test is applied to the null data. Once the p-values are obtained, we compute the proportion of significant scenarios and compare it to what is expected under H_0 (*i.e.* number of iterations times the number of tested scenarios times the threshold). We choose the conventional threshold of 5%.

This procedure estimates the false positive rate (type-I error rate *i.e.* proportion of an incorrect rejection of the null hypothesis amongst all tested hypotheses) and allows validating the tests without thorough exploration of the parameter space (i.e., the approach explained in 2.4.1).

2.4.3. H1: fMRI data. In order to assess the sensitivity of different tests, we apply all tests to the same data and a wide range of exemplar-discriminability test scenarios. To do this, fMRI data from six regions of interest (i.e., V1, V2, V3, LOC, FFA, and PPA) of 17 subjects

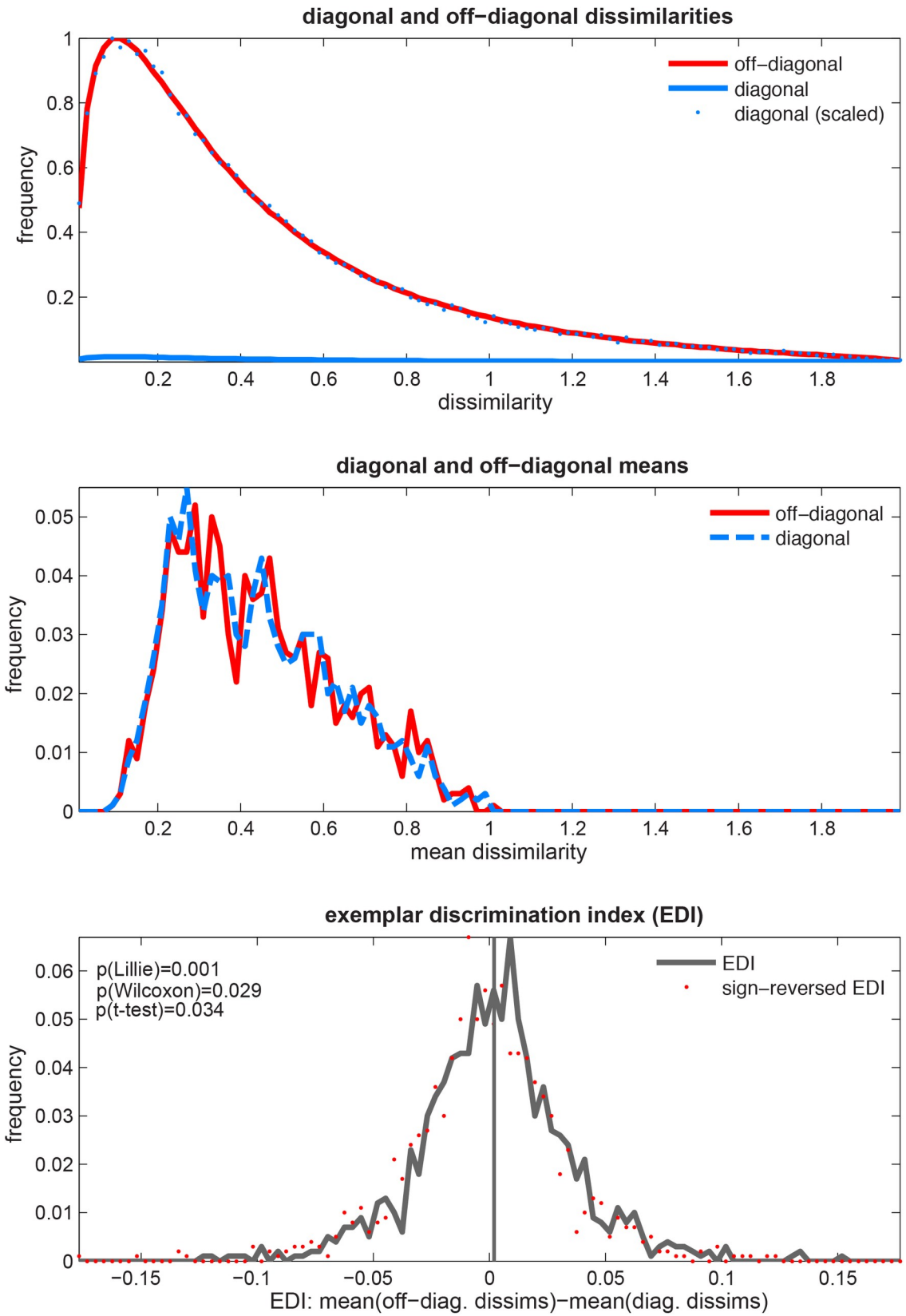

Fig 6.

https://doi.org/10.1371/journal.pone.0232551.g006

were considered. In addition, we consider discriminability of different subsets of experimental conditions. Fig 5 shows the regions of interest and the different stimulus sets. Pattern dissimilarities were computed based on the beta coefficients from the GLM. The design matrix consisted of one regressor per exemplar and six motion parameters.

Having six ROIs and seven discrimination sets (42 scenarios in total) includes a range of effects. For a given threshold, the number of significant scenarios is an indicator of the power. A test that is more powerful will give more significant cases compared to a less powerful test. If an effect is very strong, all tests would most likely detect it (resulting in significantly small p-value) but a weak effect will only be detected by a sensitive test. Thus, the comparison is fair since different tests are applied to the same data and also reasonably general since a range of effects are considered.

Applying all tests to the same dataset would not allow statistical comparison of the power of different tests. For that, we need to estimate the "variability" in the power of tests as well. To do that, we employ subject bootstrapping (resampling subjects with replacement), repeating the same procedure for a large number of subject replacements. At each bootstrap iteration, tests are applied to the bootstrapped group-data and the number of significant scenarios are counted. The standard deviation of the bootstrapping distribution would be an estimate of the standard error of the mean for the actual data [36]. We obtain p-values for each pair of tests by computing the proportion of cases where the number of significant cases is different in the two tests (*e.g.* for two tests, $test_1$ and $test_2$, the p-value for the null hypothesis that the power of $test_1$ is greater than the power of $test_2$ is the proportion of bootstrapping iterations in which the number of significant cases reported by $test_1$ is less than or equal to the number of significant cases reported by $test_2$).

3. Results

3.1. Simulation results: All tests are empirically valid

The simulations enabled us to test hypotheses about the EDI distribution under H_0. In particular, we explored a wide range of parameter settings (*e.g.* number of exemplars, *etc.*). For each setting, we assessed the validity of the *t* testing approach by testing whether the null distribution conforms to the *t* test distributional assumptions and if the *t* test protects against false positives at a reasonable rate. Furthermore, we simulated H_0 using real fMRI data and obtained estimates of the false-positive rate for all exemplar discriminability tests (explained in section 2.3). In both cases (simulating H_0 using real or simulated data), if the tests are valid, the false-positive rates should not be different from what is expected to be significant by chance (the number of false-positives depends on the threshold and the number of tests, e.g. when we apply a threshold of 5% to 100 tests under H_0, we expect 5 tests to be significant).

3.1.1. The *t* test assumptions are not met: The EDI is zero-mean, but not Gaussian under H_0. Although the EDI is not exactly symmetrically distributed about 0 under H_0, in practice it comes very close to a symmetric distribution about 0, for three reasons. (1) Though the representational distances are positively biased and have an asymmetric distribution, their distribution becomes more and more symmetrical as the dimensionality of the patterns (i.e. the number of response channels, e.g. voxels) increases. Even for very small numbers of response channels (e.g. five) the distances approximate a symmetrical distribution under H_0. (2) When dissimilarities are averaged to obtain diagonal and off-diagonal means, the distribution of each of these means even more closely approximates symmetry as it becomes more Gaussian (according to the central limit theorem). (3) The variances of the diagonal and off-diagonal means are clearly different, but this difference is smaller than one might intuitively expect. The diagonal elements are independent, so their average has a variance reduced by

factor $\sqrt{N}$. The off-diagonal elements are dependent and although N^2-N of them are averaged, the reduction in the variance is much smaller than factor $\sqrt{N^2 - N}$.

For these reasons, it is very difficult indeed to find a scenario by simulation where the EDI is not approximately symmetrically distributed about 0 under H_0. Fig 6 gives an example of a simulated extreme scenario, which we selected to illustrate the theoretical violations of the normality and symmetry assumptions. The upper two panels illustrate the approximate symmetry of the dissimilarities and their diagonal and off-diagonal means. The lower panel shows the EDI distribution, which is also close to, but not exactly, zero-mean symmetric. The EDI in this selected simulated null scenario is slightly but significantly non-Gaussian (Lilliefors test) and the *t* test and Wilcoxon signed-rank test are also significant.

3.1.2. The *t* test is empirically valid because of its robustness to violations of its assumptions. In the previous section we showed that the EDI is not strictly Gaussian under H_0. Here, we used simulations to see how often these violations occur. To this end, sdRDMs were computed under the null hypothesis for a large number of simulated subjects. Simulations were carried out independently for every point of the 5-dimensional parameter space. The parameters were the number of exemplars, number of response channels (*e.g.* voxels) and three others characterizing the multivariate response space (see Fig 4 and section 2.4.1). At each point of the parameter space, we tested if the EDI null-distribution was Gaussian and if it was zero-centered (the two main assumptions of the *t* test of EDIs). Additionally, for every point, we estimated the false positive rate of the *t* test. Estimating the false positive rate at each point in the parameter space was carried out by independently repeating the whole simulation 1,000 times and then obtaining the false positive rates from the proportion of cases where the *t* test is significant for a specified threshold.

Fig 7 shows the results for the Lilliefors test. Each panel corresponds to one of the parameters of the parameter space (5 in total, see section 2.4.1). Each bar graph gives the marginal histograms for the frequency of violations of the Lilliefors test for different levels of a parameter. Frequency of violations for the other two tests, i.e. *t* test and signed-rank test were close to zero for the different levels of all parameters (after controlling FDR at 5%) and not displayed. As expected, the null distribution is very rarely centered at a point different from zero.

Importantly, in accordance with the theoretical argument for the non-Gaussianity of the EDI, cases in which the EDI null distribution was significantly non-Gaussian were not infrequent. The validity of the assumptions also seemed to depend on the parameter level (e.g. fewer response channels are more likely to give rise to a non-Gaussian EDI null distribution). Interestingly, despite these violations of assumptions, the false-positive rates of the *t* test were not unacceptable in any of the tested scenarios (corrected for multiple comparisons). These seemingly contradictory results, i.e. theoretical violations of Gaussianity and acceptable false-positive rates, can be explained by the fact that the *t* test is robust to violations of its assumptions. Overall, these results suggest that the assumptions are violated in some cases, but the violations are small and do not significantly inflate the false positives rate of the *t* test.

3.1.3. All tests of exemplar information protect against false positives at the expected rate. Our simulations showed that the one-sided *t* test is a valid approach to test EDIs. Section 2.3 listed a variety of EDI tests including the commonly used *t* test. In this section we estimated the false positive rate (type-I error rate) of all exemplar information tests. To this end, we applied the tests to many instantiations of group data under H_0. Each instantiation was obtained by shuffling fMRI data (as explained in 2.4.2). Fig 8 shows the false-positives rates of the different tests.

As the results show, all tests have an acceptable false positive error rate: at a significance threshold of 5%, about 5% of the tests give significant results.

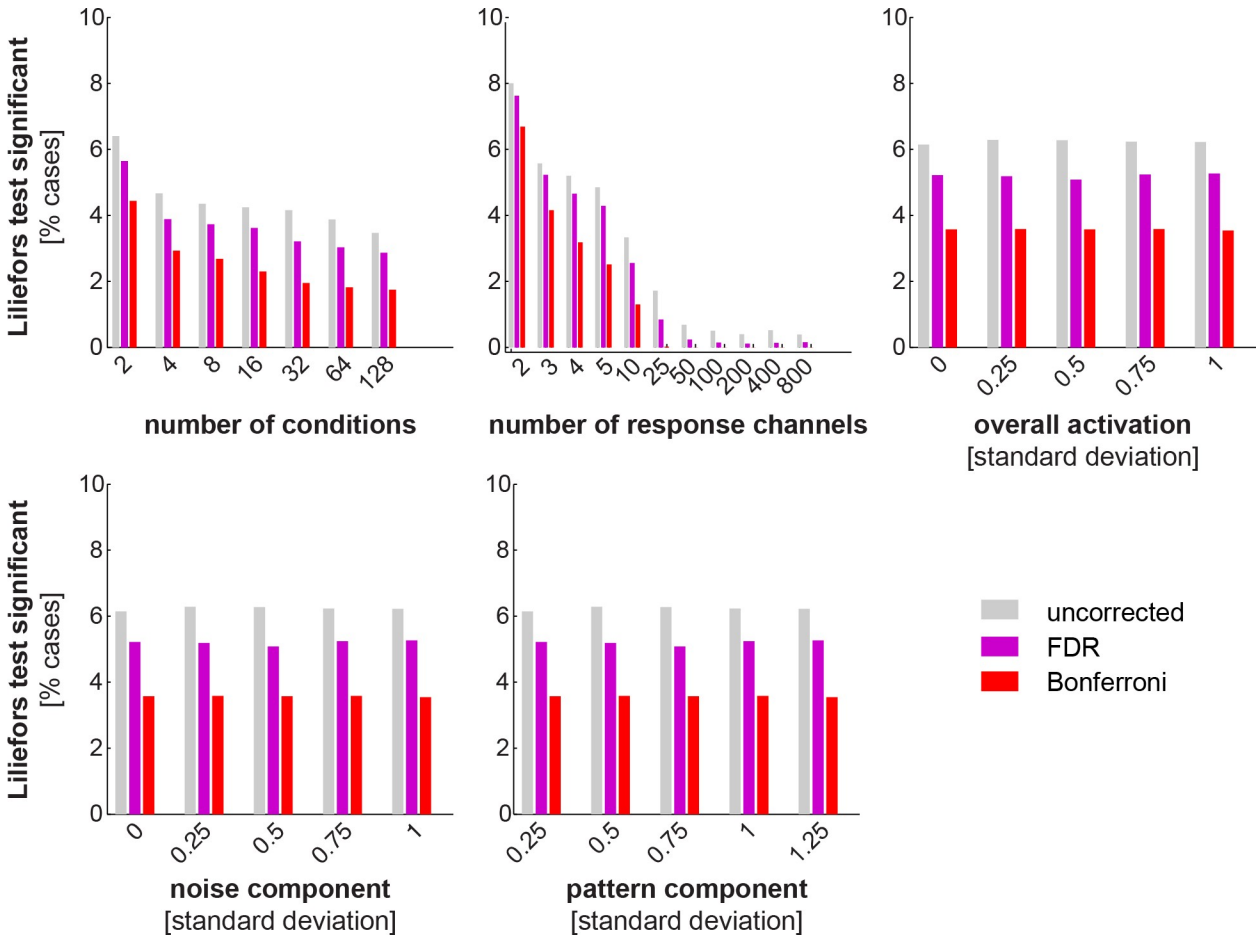

Fig 7.

https://doi.org/10.1371/journal.pone.0232551.g007

Simulating the null hypothesis using real data is more realistic than using simulations, as it incorporates all the complexities of measured data. In simulations, it can be impractical to model voxel dependencies or the extent to which response patterns change from one session to the other (different data splits). Using real data inherits all those dependencies and allows answering the same questions. However, one disadvantage of real data is that parameters of interest cannot be studied as principled as in simulations and that the conclusions may not hold for other types of neuroimaging data. In this case the consistency between the results from the simulated and real data makes us confident that all the proposed tests are valid and can be used to test exemplar information from brain measurements.

It must be noted that having a reasonable false positive rate is a necessary criterion for the validity of a test and if any test gives greater false positive rate than what is expected by chance, the test would not be considered further.

3.1.4. The EDI and linear decoders are sensitive to variance differences between conditions, whereas crossvalidated distance estimators are not. The multivariate response to each experimental condition can be treated as a sample from a high-dimensional probability distribution. The condition-related distributional differences may be of neuroscientific interest. For example, exemplar information may be present in higher-order activity-pattern statistics. Testing for mean differences can miss those effects. If two condition-related pattern

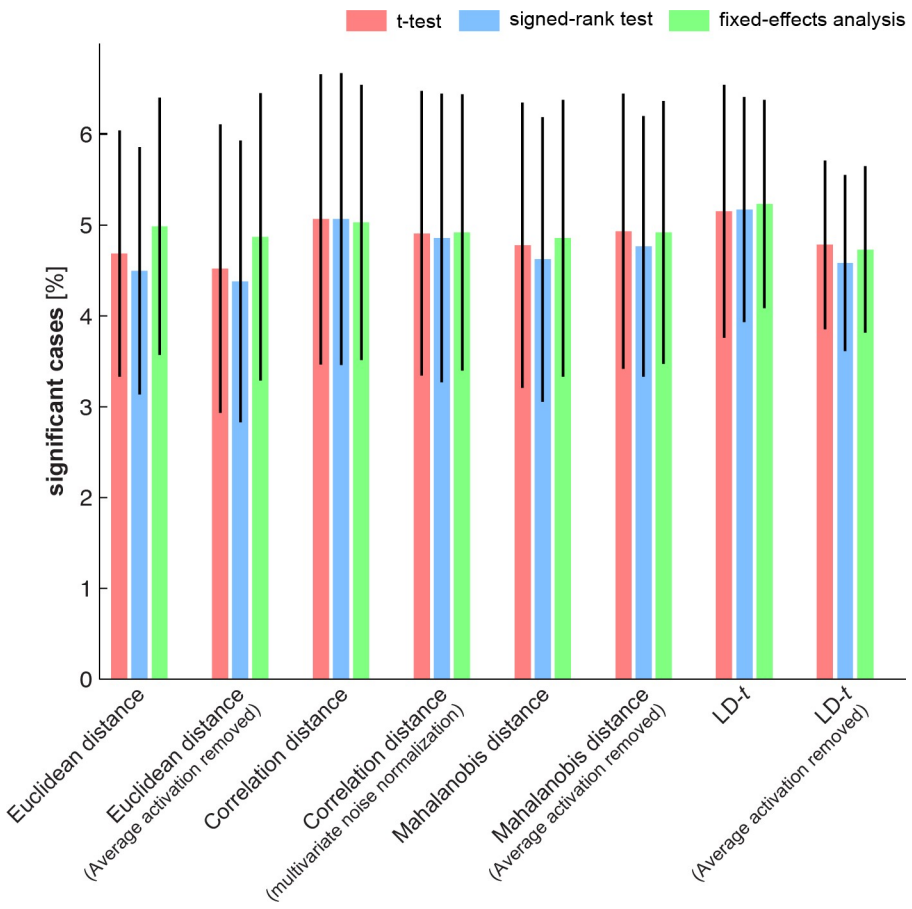

Fig 8.

https://doi.org/10.1371/journal.pone.0232551.g008

distributions have identical means and different variances (Fig 9), then there is mutual information between the response pattern and the experimental condition, which the brain might exploit.

Fig 9A illustrates that the EDI is sensitive to differences in the condition-related variances when the pattern means are identical. We consider two conditions A and B sharing the same mean pattern. The pattern-estimate distribution for A has small variance and that for B has large variance. To complement the visual intuition provided by Fig 9 with a simple numerical example, consider the extreme case where the variance is zero for condition A. The Euclidean distance between samples **a** (from A) and **b** (from B) is the root mean square (across voxels) of the difference **b**-**a**, so we can use the standard deviations (normalized root mean squares) of the distances with equivalent results (down to the proportionality factor). The expected distance between pattern estimates from A and B here is $\sqrt{(var(A)+var(B))}=\sqrt{0+1}=1$. The expected distance between different pattern estimates from A is $\sqrt{var(A)+var(A)}=0$. The expected distance between different pattern estimates from B is $\sqrt{var(B)+var(B)}=\sqrt{2}$ So the expected EDI is $1-(0+\sqrt{2})/2$ and larger than zero.

Fig 9B illustrates that a linear decoder is also sensitive to differences in the condition-related variances [11]. Placing the decision boundary to one side of the smaller-variance distribution can yield above-chance decoding accuracy. Note that our illustration here uses isotropic

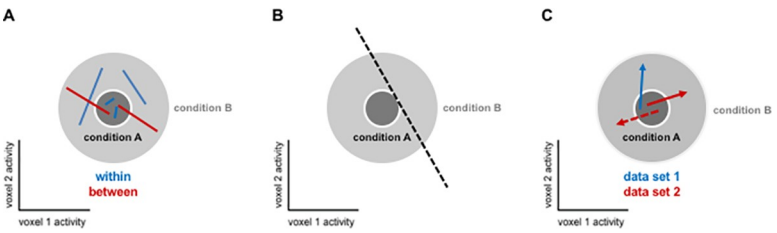

Fig 9.

https://doi.org/10.1371/journal.pone.0232551.g009

distributions. However, similar results obtain for nonisotropic distributions, and spatial whitening does not resolve this issue.

Although the brain might exploit variance differences using linear or nonlinear readout mechanisms, these are more difficult to interpret than mean differences. Moreover, in designs where conditions are not balanced (e.g. different numbers of repetitions for different conditions), the pattern estimates will differ in their variance even if there is no information about the exemplars in the responses at all. In this scenario, the variance differences do not reflect neuronal information that the brain might exploit, but arise as an artefact of the experimental design and analysis (with the researcher injecting the information about the conditions by

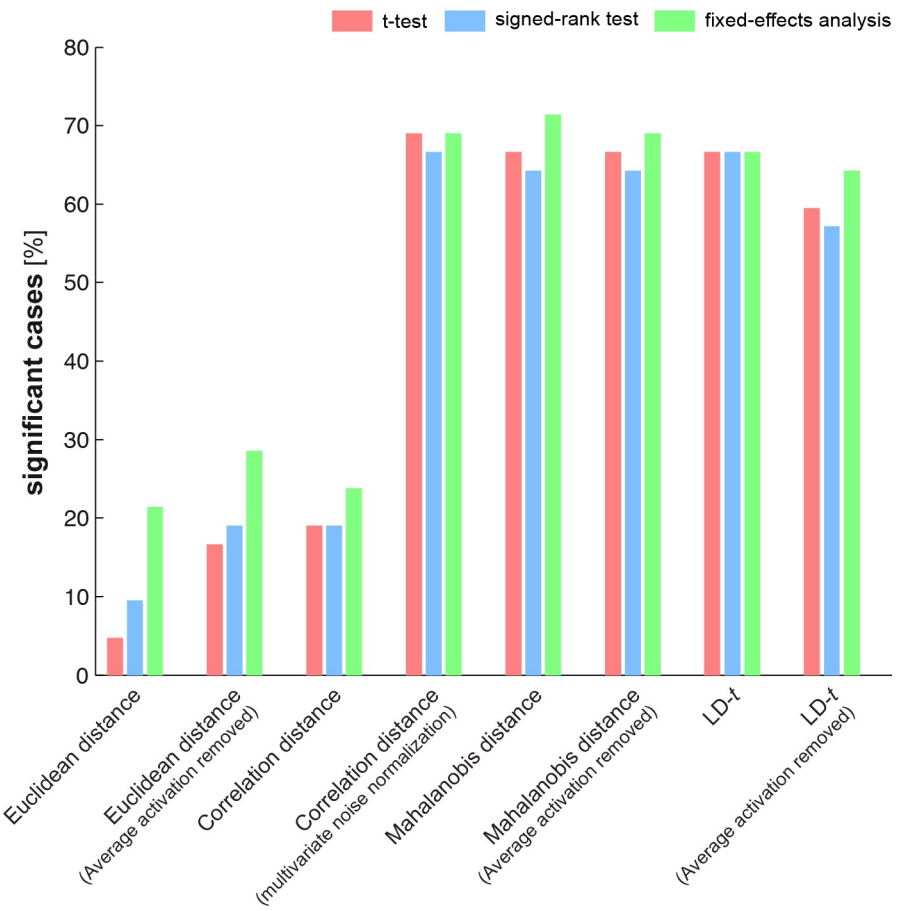

Fig 10.

https://doi.org/10.1371/journal.pone.0232551.g010

using the design matrix). The EDI and linear decoders are therefore not appropriate measures of pattern discriminability for unbalanced designs.

We may prefer more conservative measures of pattern discriminability, which are sensitive to differences in pattern mean, but not to differences in pattern variance. Crossvalidated distance estimators, such as the LDC (also known as the crossnobis estimator) or the LD-*t* [16, 28, 29] provide sensitivity to mean differences, but not to variance differences, as illustrated in Fig 9C. Crossvalidated distance estimators do not use a threshold but estimate the distance as a mean of inner products. As long as the condition-related distributions are point-symmetric, these estimators are zero in expectation (unbiased) when the pattern means are identical. This still holds when the distributions differ (in variance and/or shape) between the two conditions. In particular, the distributions could be nonisotropic Gaussian distributions with different covariance matrices. Having more data (e.g. more independent measurements) would bring the estimates closer to the true values and can always improve the precision.

3.1.5. EDI versus exemplar decoding accuracy. In addition to obtaining an EDI, split-data RDMs could also be used to obtain an estimate of exemplar decoding accuracy. A minimum-distance classifier would be able to successfully decode exemplars if the nearest neighbour of the response to each exemplar is also from the same exemplar (i.e. its replicate). Therefore, for two exemplars and two datasets, each within-exemplar distance needs to be less than the two between-exemplar distances (4 inequalities) and the decoding accuracy would be the percentage of inequalities that are satisfied. To obtain the exemplar decoding accuracy for a set of exemplars, each diagonal element of a sdRDM (N elements) would be compared to each corresponding off-diagonal element (2^*(N-1) elements) and the total number of satisfied inequalities would be counted. Normalising this count by the total number of inequalities, i.e. 2N(N-1) gives an estimate of the exemplar decoding accuracy for an exemplar set. The exemplar decoding accuracy is expected to be less sensitive than the EDI because of the loss of information incurred by the counting [29]. Moreover, decoding accuracy saturates at 100%, whereas the EDI continuously measures the separation of the exemplars. For these reasons, continuous non-saturating measures, including the EDI and pair-averaged crossvalidated distance estimates, appear preferable to exemplar decoding accuracy.

Another theoretical difference between EDI and decoding accuracy (obtained from linear decoders) is that EDI does not only capture linear information. For example, the EDI can be positive in cases where a linear classifier cannot successfully decode two exemplars. Similar to linear decoders and unlike EDI, average LDt or LDC also quantify the linear separability of representations.

3.1.6. Interpretational caveat of random-effects analysis for EDI. Allefeld et al. [37] have argued that testing classification accuracies to chance level through the random-effects analysis implemented by a t-test does not provide population inference. The reasoning is that the true level of the measures can never be below chance level. The same argument can be applied to EDI values. The true EDI, i.e. the EDI estimated from patterns without noise, will always be positive: the diagonals would all be zero and off-diagonal entries are also positive. Therefore, the null hypothesis that there is no exemplar information would translate to the average EDI being zero. Now, the average EDI for the population can only be zero if the EDI is zero in every single subject, effectively the null hypothesis for the fixed-effect test of the average EDI. Therefore, the random and fixed effects would converge in this case. The random-effect tests we suggest takes into account the between-subject variability and would be more conservative than the tests that treat subjects as fixed effect. Altogether, the across-subject EDI tests are valid tests that are subject to an interpretational caveat.

One way to circumvent that would be to convert EDI estimates to classification accuracies and test for information prevalence [37]. However, as noted earlier, converting EDI to

classification accuracies brings the undesirable property that the measure will saturate and cannot get any larger for fully discriminable patterns [29].

3.1.7. Separating differences in overall activations and differences in patterns. Generally, EDI tests exploit all the information present in regional responses. This also includes overall activation differences. In section 2.2.3 we suggest various ways of removing the overall activation effect when computing discriminability measures. Another way to quantify the contribution of regional mean activation effects on exemplar discriminability would be to test if the exemplars could be decoded on the basis of the global responses alone. This could be achieved by testing EDIs from activation-based *sdRDM*s. In that case entries of the *sdRDM* would be the absolute value of the differences in the voxel-averaged regional response for same or different exemplars in two datasets. A significant univariate EDI would imply that the exemplar discriminability is at least partly driven by overall activation differences. This contribution could be removed by regressing out the activation-based *sdRDM* from the original *sdRDM*.

3.2. Real-fMRI-data results: Significantly greater sensitivity through multivariate noise normalization

To complete our assessment of exemplar information tests, we compared their sensitivity, i.e. the power of the tests to detect an effect when present. One way to proceed would be to simulate data with known ground truth (i.e. exemplar discriminabilities), apply the tests to the simulated data and compare them (or their ranks) in terms of their significances. However, this approach is both computationally expensive and unrealistic. In the parameter space, each parameter can span an enormous range of values. Therefore, it would be impractical to carry out a comparison for each possible combination of parameters. Moreover, as mentioned earlier, voxel dependencies and appropriate levels of replicability observed in real fMRI data are not known a priori. For these reasons we compared the tests by applying all of them to the same dataset. The dataset contained six brain regions and seven exemplar subsets (see section 2.4.3 and Fig 5). Therefore, the total number of significant scenarios for a given test could vary from 0 to 42. Note that in order to make the test comparisons valid and fair, the same significance threshold was applied in all tests. Fig 10 gives an estimate of the sensitivity of different tests based on the real fMRI dataset.

The results show a striking effect of multivariate noise normalization (Eq 1) in the sensitivity of EDI tests. All tests that are applied to data after multivariate noise normalization yielded a greater proportion of significant cases. Statistical comparison of the tests was carried out by bootstrapping subjects and applying the tests to the group data at each bootstrapping iteration. Obtaining p-values for comparing pairs of tests and controlling the false discovery rate at 1%, we concluded that the difference between the two groups of tests was significant. In other words, summary statistics that were obtained from data after accounting for the noise covariance between voxels had significantly greater power in detecting exemplar effects.

4. Discussion

This article explores different approaches to testing subtle within-category effects in brain representations (motivated by the theoretically-invalid t-test that was common in the literature).

4.1. Testing for exemplar information in condition-rich designs is important for understanding brain representations

Understanding brain representations asks for characterising the capabilities of the representations. One important characteristic would be to allow discrimination between different

members of a category i.e. category exemplars. One could use a classifier and test for exemplar information by calculating the performance of a classifier that is trained to distinguish between different exemplars (e.g. employing a support vector machine, [38, 39]).

In contrast to the classification-based approach that requires having many repetitions of each exemplar, we aim to extract exemplar information for condition-rich designs where many exemplars are used with fewer repetitions for each. Given the limited time for measuring brain representations, there would be a trade-off between the number of tested exemplars and number of repetitions of each exemplar. Resolving this trade-off would allow higher inferential power and a stronger claim about the representational structure of a brain region. For example, if we could have more exemplars *i.e.* richer sampling of the space of exemplars, inference would be based on a larger group and that would reduce the sampling error. *Condition-rich decoding* refers to quantifying and assessing exemplar information in cases where discriminability of many experimental conditions (e.g. category exemplars) are tested.

Condition-rich decoding from distributed brain representations is possible using a simple and intuitive idea: if (on average) repetitions of the same exemplar produce less pattern-change than that produced by changing the exemplar, there is exemplar information in brain representations. Since this approach averages the effects across many pairs of exemplars, presence of exemplar information does not imply that any two pairs would be discriminable but that on average there is enough information to discriminate between exemplars on the basis of brain representations. Furthermore, in order to understand what exemplar distinctions are *driving* a significant EDI in an ROI, one can consider follow-up tests within the same framework. For example, in the extreme scenario, distinctions between any pair of conditions can be tested at the single-subject level or the group-level by testing the LD-t [28] or the crossnobis [29]. Additionally, one can consider exploring the effect for each individual subject by investigating the EDI significance at the single-subject level. Therefore, although a significant EDI does not preclude the possibility that the effect might be driven by some set of conditions or not present in all subjects, the tools we provide here allows clarification in full depth.

4.2. The conventional *t* test approach is theoretically problematic but practically acceptable

The conventional approach for condition-rich decoding uses *t* test on the EDIs. The *t* test is a parametric test and, for a given sample, tests if a parameter of the sample distribution, its mean, is different from zero (i.e. the mean of the Gaussian distribution obtained under the null hypothesis). Using parametric statistics can be more powerful when the required assumptions are satisfied. However, if the requirements are not met, test results are not interpretable [40]. For reasons that we explain in the paper, the null distribution of the EDI is likely to violate the required assumptions of the *t* test (e.g. violation of the Gaussianity assumption; see section 3.1.1). Therefore, we speculated that *t* test would be problematic. Using simulations, we first appreciate this concern by observing cases where applying a *t* test is wrong. This was then followed by a systematic exploration of the space of possible parameters (characterizing the response space in a wide range of settings) and showing that violations of assumptions are indeed frequent but to the extent that could mostly be tolerated by the test (although the *t* test presumes distributional properties, it is tolerant to violations of the assumptions).

4.3. Testing exemplar information at the single-subject level or group-level with subject as fixed effect is possible using novel randomization tests

Randomization tests, and more generally non-parametric statistics, are becoming more and more popular in the univariate analysis of neuroimaging data [40]. We can also use

randomization tests and non-parametric statistics in the context of testing for information in brain response-patterns. For example, randomization tests have been proposed before to test for the similarity (i.e. correlation) of two RDMs [6, 28]. Similarly, in the case of testing EDIs, randomization techniques could be employed to design non-parametric tests of exemplar information.

Hitherto, it has not been possible to test for exemplar information at the single-subject level or to do group-level analysis with subject as the fixed effect. Here we introduce tests for performing fixed effect analysis. To our knowledge this is the first attempt to fixed effect analysis of exemplar information using pattern-information analysis especially in the context of condition-rich designs. Our tests are based on randomization methods that estimate the distribution of the statistic under H_0. For group-level analysis, these tests are more sensitive since they ignore the between-subject variability.

4.4. Tests that take the covariance structure of the noise into account are significantly more powerful

In the past, only few studies have attempted to do condition-rich decoding. All those studies were based on quantifying pattern-changes by computing Pearson correlation distances and using a one-sided *t* test to test the net pattern-effect (average effect of changing exemplars minus average effect of repeating an exemplar) quantified in each subject, against zero. The *t* test is widely used in psychology and neuroscience; however, its use is less motivated for testing exemplar information in distributed patterns. The main reason against using *t* test is the theoretical concern about the distribution of the test statistic (i.e., EDI) under the null hypothesis.

Having established the validity of the standard approach, we consider different ways of testing exemplar information by exploring different tests and test statistics. In particular we introduce ways of testing exemplar information at the single-subject level or fixed effects analysis at the group level. In addition to the fixed effect tests, we also consider other tests including non-parametric alternatives to the *t* test. Those assumption-free methods can be used for testing exemplar information even in the extreme cases where *t* test is not valid.

The various test statistics considered in this paper are different due to their different ways of quantifying pattern-effects. In particular different measures could be used for computing pattern dissimilarities. Therefore, the situation is as follows: there are different ways of assessing exemplar information and the net effect could be *summarized* and test*ed* in different ways. Interestingly, we see that all the possible combinations of tests and test statistics yield reasonable false positive rates. In other words, choosing any of the proposed ways of summarizing the results and any of the tests would result in acceptable specificities. However, when looking at the sensitivities, we did not find equal levels of power for the different combinations.

By contrast, we observed that within the explored tests (i.e. *t* test, Wilcoxon signed rank test and fixed effect tests) and test statistics (i.e. EDIs based on different distance measures or average LD-*t* values), the most important factor in determining the power of the test was the way to quantify and summarize the net effect, and not the test itself. All pattern-effects that were computed after multivariate noise normalisation could be detected with a greater power than those that were not. This effect of multivariate noise normalisation afforded an almost 3-fold boost of power for the analysed dataset. Therefore, these results have clear practical implications for future studies. For example, researchers could add multivariate noise normalization to the pre-processing stages of data analysis and use the Euclidean distance—which has a simple geometric interpretation–to estimate pattern dissimilarities on the pre-processed datasets. This would have the advantage of being both simple and powerful.

We have previously presented evidence that multivariate noise normalisation renders RDMs more reliable [29]. This paper documents the benefit in the particular scenario of exemplar effects, and more specifically to both within-run and between-run distances.

4.5. Quantifying category information with the category discrimination index (CDI)

Although category information and exemplar information consider conceptually different characteristics of representational geometries, they could both be assessed using similar methods. In this paper, we focus on exemplar information. Category information could be quantified from an RDM by subtracting the average-within-category dissimilarities from the average-between-category dissimilarities *i.e*. a category discriminability index (CDI). The CDIs could then be tested in exactly the same way as EDIs. The main difference is that for CDI it is not necessary to have independent measurements of the same experimental conditions and split the data into two halves. One could use the whole dataset and compute an RDM from the full dataset. Using more data (i.e. the whole dataset as opposed to data-halves for a *sdRDM*) may likely result in more stable patterns, which would be an advantage of this approach. However, it is also possible to use *sdRDMs* to compute the CDI. This would have advantages when comparing patterns within a dataset can have a confound e.g. due to difference in temporal proximities (in that case the CDI is defined as the average between category dissimilarities minus the average within category dissimilarities from a sdRDM).

Moreover, with cross-validated dissimilarity estimates like the crossnobis or the LD-t, category information could be computed by averaging the between-category discriminabilities. Since those are unbiased estimates, a significantly positive average would imply category information. It must be noted that similar to exemplar tests, category decodability using CDI or classification accuracy would be sensitive to differences in the variances for the two categories; however, average between-category crossnobis estimates would only be sensitive to category-centroid differences. Therefore, using average classification accuracy or the CDI can give significant results in the absence of category-centroid differences in cases where the designs are not balanced (e.g. different number of exemplars for the two categories) and there are variance differences due to the unbalanced design.

On a similar vein to testing EDIs, one can also subtract the average within-category dissimilarities from average between-category dissimilarities for the cross-validated dissimilarities. In this case a significant result would imply category *clustering*, which is a stricter characteristic compared to linear category decodability. For example, two point-clouds can be linearly separable but not form clusters.

Another approach to testing category information is to test the rank correlation of RDMs with a categorical RDM that assumes equal and smaller values (*e.g*. zero) for within-category dissimilarities and larger values (*e.g*. one) for between-category dissimilarities [41, 42]. We have recently proposed using Kendal's tau-a for RDM correlations in such cases where tied ranks are predicted for either between- or within-category dissimilarities [28]. In cases where there are only two categories (e.g. a binary model RDM), testing rank correlations and linear correlations would be the same. However, when there are more than two levels in the discrete model RDM, testing rank correlations would be different and more lenient than linear ones. (note that for a binary model RDM the CDI for one dataset is proportional to the linear correlation of the model RDM with the data RDM)

An alternative approach is to fit categorical RDMs to brain RDMs using linear regression and test the regression coefficients against zero [42, 43]. Similar to cross-validated RDMs, category information could be also tested by testing the average-between-category classification

accuracies (Cichy et al. [44], although see [29], for a comparison of continuous measures like crossnobis versus discretised measures like classification accuracies for estimating dissimilarities).

4.6. Relation to model testing

Representational similarity analysis enables testing computational models of information processing in the brain. This is achieved by comparing the predicted representational geometries of models with observed geometries of the brain. Computational models can also imply exemplar information but only test for a particular representational geometry, i.e. particular configuration of exemplars in the response space. Therefore testing a model is a more focused test which would be more powerful if the hypothesis is true.

Note that the EDI is maximally general in that it is sensitive to any differences between conditions. It is highly powerful when differences exist for many of the pairs, but in principle it can detect a difference between just one pair of exemplars. Note also that the EDI can detect equidistant representational geometries, to which tests of RDM models using correlation coefficients are not sensitive (since there is no variability in the predicted distances that could be explained by a model RDM).

Additionally, one could use sdRDMs (Fig 1) to test computational models. The model prediction would be a symmetric matrix that predicts zeros along the diagonal entries. This would enable us to test the model predictions of the dissimilarities at the level of a ratio scale, not just an interval scale. It would increase the power when the model predicts very similar distances among the representational patterns. (Consider the extreme case where the model predicts equidistant patterns. In this scenario the sdRDM is required, and the model test would be equivalent to the EDI if the Pearson correlation was used). More generally, quantifying brain representations by a sdRDM would allow modelling the condition-wise differences in noise levels.

Some researchers have recommended using between-run distances for RSA [45]. They have shown that in cases where there are unavoidable structures to trial order and between-subject randomisation cannot be afforded, valid distance estimates can be obtained only when activity patterns are compared in two independent datasets (e.g. different fMRI runs). Cross-run RDMs can be computed through comparing pattern estimates from one run to the average of all other runs (ideally optimally combining the patterns would be preferred to averaging them). The importance of presentation-sequence randomisation is not limited to single-trial RDMs. Also in RDMs for which the activity patterns are based on multiple repetitions of an item, for instance, if a condition is always presented before another condition in the first run and in the reverse order in the other run, this can increase the odds of getting a negative crossnobis estimate due to the effects of the scanner drift. Therefore, in cases where the presentation sequences are not randomised across independent measurements (e.g. fMRI runs) and subjects, using between-run distances after multivariate noise normalisation might be a good compromise. Surely, one could ensure that the design is balanced and use crossnobis for computing RDMs.

4.7. Choosing the appropriate summary statistic for condition-rich decoding

In this manuscript we have explored a range of summary statistics and tests for exemplar discriminability. Table 1 summarizes the characteristics of the summary statistics. We recommend using the stimulus-pair averaged crossnobis or LDt for testing mean differences and

Table 1. An overall view of different methods for testing exemplar discriminability.

test statistic	multivar. noise model	sensitive to	H_0	inference procedure	inference scope	Validity (specificity)	Power (sensitivity)
exemplar discriminability index (EDI): mean of between-exemplar distance minus mean of within-exemplar distances	no	differences in pattern distributions (incl. mean and variance)	exemplar response-patterns drawn from the same distribution	one-sided t test across subjects	subject population	usually acceptable (despite violations of assumptions)	bad
				one-sided signed-rank test across subjects		good	bad
				exemplar-label randomization	subject sample	good	good
	yes			one-sided t test across subjects	subject population	usually acceptable (despite violations of assumptions)	very good
				one-sided signed-rank test across subjects		good	very good
				condition-label randomization	subject sample		excellent
average of pairwise crossvalidated discriminabilities (crossnobis, LD-*t*)		differences in pattern means	exemplar response-patterns drawn from distributions centered on the same mean pattern	one-sided t test across subjects	subject population		very good
				one-sided signed-rank test across subjects			very good
				exemplar-label randomization (different exemplar labels between training and test sets)	subject sample		excellent
crossvalidated MANOVA (pattern distinctness)				t test, signed-rank, permutation	subject sample or population		very good, excellent

https://doi.org/10.1371/journal.pone.0232551.t001

EDIs based on multivariate noise normalisation for testing effects in mean and/or variance differences.

5. Conclusions

When independent measurements of the same set of exemplars are available, condition-rich decoding can be carried in different ways. Despite the theoretical concerns against applying the *t* test to the EDIs, we validated the use of the *t* test, which is the established approach for testing exemplar information. Furthermore, we explored different ways of summarizing and testing exemplar discriminabilities and introduced a novel randomization test that is assumption-free and also allows testing EDIs at the single-subject level or group level analysis with subject as fixed effect. Comparing the different methods, we conclude that it is mostly critical to compute the EDI after applying multivariate noise normalization to the response patterns. Multivariate noise normalization considers the noise variance-covariance structure of the data and enables a more reliable estimation of RDMs. This means that computing the EDIs after multivariate noise normalization or using the average LD-*t* or crossnobis will likely result in higher sensitivity to detect exemplar information in brain representations. We also explain the difference between exemplar (and category) tests that are based on average crossnobis or LD-t and those that are based on EDI or average decoding accuracies. Average crossnobis or LD-t is only sensitive to differences in pattern mean whereas EDI and decoding accuracies are also sensitive to variance differences.

Author Contributions

Conceptualization: Hamed Nili, Nikolaus Kriegeskorte.

Data curation: Arjen Alink.

Formal analysis: Hamed Nili, Alexander Walther, Nikolaus Kriegeskorte.

Funding acquisition: Nikolaus Kriegeskorte.

Investigation: Hamed Nili, Alexander Walther, Arjen Alink, Nikolaus Kriegeskorte.

Methodology: Hamed Nili, Alexander Walther.

Project administration: Nikolaus Kriegeskorte.

Software: Hamed Nili, Alexander Walther, Nikolaus Kriegeskorte.

Supervision: Nikolaus Kriegeskorte.

Validation: Hamed Nili.

Visualization: Hamed Nili, Alexander Walther.

Writing – original draft: Hamed Nili, Alexander Walther, Arjen Alink, Nikolaus Kriegeskorte.

Writing – review & editing: Hamed Nili.

References

1. Haxby J. V., Gobbini M. I., Furey M. L., Ishai A., Schouten J. L., & Pietrini P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430. https://doi.org/10.1126/science.1063736 PMID: 11577229
2. Hung C.P., Kreiman G., Poggio T., DiCarlo J.J.,2005. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science 310: 863–866. https://doi.org/10.1126/science.1117593 PMID: 16272124
3. Kamitani Y., & Tong F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679–685. https://doi.org/10.1038/nn1444 PMID: 15852014
4. Kriegeskorte N., Goebel R., & Bandettini P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America, 103(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103 PMID: 16537458
5. Kiani R., Esteky H., Mirpour K., Tanaka K., 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of Neurophysiology 97, 4296–4309. https://doi.org/10.1152/jn.00024.2007 PMID: 17428910
6. Kriegeskorte N., Mur M., Ruff D. A., Kiani R., Bodurka J., Esteky H., et al. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126–1141.
7. Kriegeskorte N., & Kreiman G. (Eds.). (2012). Visual population codes: toward a common multivariate framework for cell recording and functional imaging. MIT press.
8. Kriegeskorte N., & Kievit R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007 PMID: 23876494
9. Mur M., Bandettini P. A., & Kriegeskorte N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide.*Social cognitive and affective neuroscience*, 4(1), 101–109. https://doi.org/10.1093/scan/nsn044 PMID: 19151374
10. Norman K. A., Polyn S. M., Detre G. J., & Haxby J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005 PMID: 16899397
11. Hebart M. N., & Baker C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005 PMID: 28782682

12. Kanwisher N., McDermott J., & Chun M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997 PMID: 9151747

13. Kay K. N., Naselaris T., Prenger R. J., & Gallant J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355. https://doi.org/10.1038/nature06713 PMID: 18322462

14. Mitchell T. M., Shinkareva S. V., Carlson A., Chang K. M., Malave V. L., Mason R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880), 1191–1195. https://doi.org/10.1126/science.1152876 PMID: 18511683

15. Anzellotti S., Fairhall S. L., & Caramazza A. (2013). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, bht046.

16. Kriegeskorte N., Formisano E., Sorger B., & Goebel R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104 (51), 20600–20605.

17. Nestor A., Plaut D. C., & Behrmann M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24), 9998–10003.

18. Chan A. W., Kravitz D. J., Truong S., Arizpe J., & Baker C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature neuroscience*, 13(4), 417–418. https://doi.org/10.1038/nn.2502 PMID: 20208528

19. Kravitz D. J., Kriegeskorte N., & Baker C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*, 20(12), 2916–2925. https://doi.org/10.1093/cercor/bhq042 PMID: 20351021

20. Lee S. H., Kravitz D. J., & Baker C. I. (2012). Disentangling visual imagery and perception of real-world objects. Neuroimage, 59(4), 4064–4073. https://doi.org/10.1016/j.neuroimage.2011.10.055 PMID: 22040738

21. Liu N., Kriegeskorte N., Mur M., Hadj-Bouziane F., Luh W. M., Tootell R. B., et al. (2013). Intrinsic structure of visual exemplar and category representations in macaque brain. *The Journal of Neuroscience*, 33(28), 11346–11360. https://doi.org/10.1523/JNEUROSCI.4180-12.2013 PMID: 23843508

22. Luyckx F., Nili H., Spitzer B., & Summerfield C. (2019). Neural structure mapping in human probabilistic reward learning. *eLife*, 8, e42816. https://doi.org/10.7554/eLife.42816 PMID: 30843789

23. Sayres R., & Grill-Spector K. (2008). Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of Neurophysiology*, 100(1), 249. https://doi.org/10.1152/jn.01383.2007 PMID: 18463186

24. Schwarzlose R. F., Swisher J. D., Dang S., & Kanwisher N. (2008). The distribution of category and location information across object-selective regions in human visual cortex. *Proceedings of the National Academy of Sciences*, 105(11), 4447–4452.

25. Spitzer B., Waschke L., & Summerfield C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, 1(8), 0145.

26. Alink A., Walther A., Krugliak A., van den Bosch J. J., & Kriegeskorte N. (2015). Mind the drift-improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. bioRxiv, 032391.

27. Henriksson L., Khaligh-Razavi S. M., Kay K., & Kriegeskorte N. (2014). Intrinsic cortical dynamics dominate population responses to natural images across human visual cortex. *bioRxiv*, 008961.

28. Nili H., Wingfield C., Walther A., Su L., Marslen-Wilson W., & Kriegeskorte N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553 PMID: 24743308

29. Walther A., Nili H., Ejaz N., Alink A., Kriegeskorte N., & Diedrichsen J. (2015). Reliability of dissimilarity measures for multi-voxel pattern analysis.*NeuroImage*, 137, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012 PMID: 26707889

30. Kriegeskorte N., Mur M., & Bandettini P. (2008a). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2.

31. Ledoit O., & Wolf M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10(5), 603–621.

32. Wilcoxon F., 1945. Individual Comparisons by Ranking Methods, Biometrics Bulletin. Vol 1, No 60, 80–83.

33. Dixon W. J., & Mood A. M. (1946). The statistical sign test. Journal of the American Statistical Association, 41(236), 557–566. https://doi.org/10.1080/01621459.1946.10501898 PMID: 20279351

34. Lilliefors H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318), 399–402.

35. Benjamini Y., & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289–300.

36. Efron B., & Tibshirani R. J. (1994). *An introduction to the bootstrap*. CRC press.

37. Allefeld C., Görgen K., & Haynes J. D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *Neuroimage*, 141, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040 PMID: 27450073

38. Burges C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121–167.

39. Misaki M., Kim Y., Bandettini P. A., & Kriegeskorte N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage, 53(1), 103–118. https://doi.org/10.1016/j.neuroimage.2010.05.051 PMID: 20580933

40. Nichols T. E., & Holmes A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1), 1–25. https://doi.org/10.1002/hbm.1058 PMID: 11747097

41. Khaligh-Razavi S. M., & Kriegeskorte N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915 PMID: 25375136

42. Mur M., Meys M., Bodurka J., Goebel R., Bandettini P. A., & Kriegeskorte N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in psychology*, 4.

43. Jozwik K. M., Kriegeskorte N., Storrs K. R., & Mur M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, 8, 1726. https://doi.org/10.3389/fpsyg.2017.01726 PMID: 29062291

44. Cichy R. M., Pantazis D., & Oliva A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, 17(3), 455–462. https://doi.org/10.1038/nn.3635 PMID: 24464044

45. Mumford J. A., Davis T., & Poldrack R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage*, 103, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026 PMID: 25241907