# Second revision:
# Linking brain-wide multivoxel activation patterns to behaviour: examples from language and math

Rajeev D. S. Raizada,[1] Feng Ming Tsao,[2] Huei-Mei Liu,[3]
Ian D. Holloway,[4] Daniel Ansari,[4] and Patricia K. Kuhl[5]

[1]Neukom Institute for Computational Science, HB 6255, Dartmouth College, Hanover NH 03755.
[2]Dept. of Psychology, National Taiwan University, Taipei 10617, Taiwan.
[3]Dept. of Special Education, National Taiwan Normal University, Taipei 10644, Taiwan.
[4]Dept. of Psychology, Univ. of Western Ontario, London, ON N6G 2K3, Canada.
[5]Inst. for Learning & Brain Sciences, Univ. of Washington, Box 357988, Seattle WA 98195.

## Abstract

A key goal of cognitive neuroscience is to find simple and direct connections between brain and behaviour. However, fMRI analysis typically involves choices between many possible options, with each choice potentially biasing any brain-behaviour correlations that emerge. Standard methods of fMRI analysis assess each voxel individually, but then face the problem of selection-bias when combining those voxels into a region-of-interest, or ROI. Multivariate pattern-based fMRI analysis methods use classifiers to analyse multiple voxels together, but can also introduce selection bias via data-reduction steps as feature selection of voxels, pre-selecting activated regions, or principal components analysis. We show here that strong brain-behaviour links can be revealed without any voxel selection or data reduction, using just plain linear regression as a classifier applied to the whole brain at once, i.e. treating each entire brain volume as a single multi-voxel pattern. The brain-behaviour correlations emerged despite the fact that the classifier was not provided with any information at all about subjects' behaviour, but instead was given only the neural data and its condition-labels. Surprisingly, more powerful classifiers such as a linear SVM and regularised logistic regression produce very similar results. We discuss some possible reasons why the very simple brain-wide linear regression model is able to find correlations with behaviour that are as strong as those obtained on the one hand from a specific ROI and on the other hand from more complex classifiers. In a manner which is unencumbered by arbitrary choices, our approach offers a method for investigating connections between brain and behaviour which is simple, rigorous and direct.

# 1 Introduction

Finding brain activation is easy; the difficult part is trying to determine its functional significance. Our best guide for this is to study behaviour: the more directly we can link brain activation to behaviour, the stronger our evidence is about what this activation might mean. However, the goal of finding links between brain and behaviour finds itself caught between seemingly conflicting aims: on the one hand, we would like to find connections which are as simple and direct as possible. On the other hand, too simple a method might be neither powerful nor sensitive enough to succeed in capturing actual brain-behaviour connections.

One difficulty which has recently received much attention is the problem of selection-bias. Typically, the relation that is sought is not between behaviour and the whole brain, but instead is between behaviour and some subregion of the brain, namely a Region-of-Interest (ROI). However, if the criterion used to define a given ROI is the same as the one being tested for in the data extracted from that ROI, then the resulting correlation test may be biased or even wholly invalid (Kriegeskorte et al., 2009; Vul et al., 2009).[1] This problem is not specific to correlation, but may potentially apply to any statistical test. Feature-selection and other forms of data-reduction, although useful when used with care, can potentially face a similar problem: if a brain-behaviour correlation emerges only when just the right set of carefully selected features is used, then its validity may be in doubt. In each of these cases, the preprocessing or selection stage inserts an extra step into the link between brain and behaviour, making any such link less direct, and more dependent upon the particular set of analysis methods which happened to be used when finding it.

It would therefore be ideal, if possible, to find brain-behaviour correlations without needing to do any selection of voxels or features, but instead by simply applying one test across the whole brain at once. However, such a test might seem infeasible: an MRI volume of the whole brain contains many voxels, only some of which would be expected to carry task-related signal. Any brain-behaviour links hidden amongst the signal-bearing voxels would, it might seem, be swamped by the inclusion of all the noise.

In the present study we show that, on the contrary, a simple classifier applied to the whole brain at once can indeed find strong brain-behaviour correlations, across two distinct data sets using very different behavioural tasks. The classifier used was perhaps the simplest of all: linear regression with the category labels +1 and -1 as the desired outputs. The resulting model classifies an input as category +1 or -1 according to whether its output is greater than or less than zero.

[1] It is worth noting that it is possible to look for brain-behaviour correlations without forming ROIs, for example by using a GLM with some regressors based on the subjects' behaviour inside the scanner. Such an analysis will produce areas of significant activation, and can be left at that point. However, it is often desirable to look inside the resulting activated regions, in order to ensure that the correlations are not due to outliers or other data pathologies such as those illustrated by Anscombe's quartet (Anscombe, 1973). At that point, the ROI selection-bias difficulties discussed by Vul et al. (2009) and Kriegeskorte et al. (2009) may once again arise.

In a standard General Linear Model (GLM) analysis, the dependent variables are written as $\mathbf{X}$ and the independent variables are written as $\mathbf{Y}$. The variable $\mathbf{X}$ is the design matrix, which provides the regressors for the linear model, and the variable $\mathbf{Y}$ is the MRI signal, which provides the data to be fitted by the GLM. The regression coefficients are standardly labeled as $\boldsymbol{\beta}$ and the residuals as $\mathbf{e}$, giving the GLM equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

In the pattern-based analysis used here, the direction of this regression is reversed. Now $\mathbf{X}$ is the MRI data, and $\mathbf{y}$ is the vector of category labels stating which experimental conditions happened at which times. The discriminant weight vector, $\mathbf{w}$, is then given by $\mathbf{w} = \mathbf{X}^{\dagger}\mathbf{y}$, where $\mathbf{X}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{X}$. The number of tunable model parameters is zero. Below, we discuss some reasons why this classifier, despite its simplicity and its having to deal with large numbers of uninformative voxels, is nonetheless able to extract behaviourally relevant information.

The two different data sets which we use come from different labs, and are drawn from different cognitive domains. The first is from a recent study of how Japanese and English subjects perceive the syllables /ra/ and /la/, in which we showed that the degree to which a classifier could separate the neural patterns elicited by those sounds predicted how well people could behaviourally tell the sounds apart, not only across groups (English vs. Japanese) but also across individuals (Raizada et al., 2010). In that study, the brain-behaviour correlation was derived using a small Heschl's gyrus ROI and an SVM classifier (see Section 4.1 for discussion of how the ROI in that previous paper was defined and used). In the present study we show that when the whole brain is analysed at once, using just plain regression, remarkably similar brain-behaviour correlations continue to hold both across groups and across individuals.

The second data set is from a study of numerical cognition (Holloway et al., 2010). The data from that study relates fMRI activation on a non-symbolic numerical "distance effect" task (Dehaene et al., 2004) to behavioural scores on standardised tests of arithmetic and language. In a distance-effect task, the subject judges which one of a pair of numbers presented together is larger; the closer together in magnitude the two numbers are, the longer the subject's reaction time, and the more activation in lateral intraparietal cortex is produced (Moyer & Landauer, 1967; Dehaene et al., 2004). The activation difference between small-distance pairs and large-distance pairs in parietal cortex has been found to relate to behavioural ability to perform numerical tasks: it is smaller in children with dyscalculia than in control children (Price et al., 2007; Mussolin et al., 2009). As is laid out in detail below, we found a whole-brain correlation with behavioural performance in this distance-effect data: the separability between neural patterns elicited by large-distance pairs and small-distance pairs was positively correlated with subjects' scores on the arithmetic tests. However, this pattern separability measure was calculated not from a parietal cortex ROI, but from the whole brain at once. Moreover, the correlation with behaviour was specific: fMRI pattern separability did not correlate with language test scores at all.

In the sections below, we describe the details of the methods used, the brain-behaviour correla-

tions which emerge, and then some possible reasons why this whole-brain regression model succeeds in revealing those correlations despite its simplicity. Finally, we discuss how this approach may offer a more direct and robust method for assigning functional meaning to brain activation than has hitherto been available.

# 2 Materials and methods

## 2.1 MRI acquisition and subject information

The full methods, subject and task details for the /ra/-/la/ data have already been described in the published literature (Raizada et al., 2010), and the full details for the numerical distance effect data are described in Holloway et al. (2010). We here summarise the key aspects of those papers' methods; for more details please refer to the original publications.

Raizada et al. (2010) scanned twenty normal adult subjects: ten were native English speakers, and ten were native speakers of Japanese. fMRI scans were carried out on a GE Signa 1.5T scanner at the University of Washington, using a standard BOLD EPI sequence: TR = 2000 ms, TE = 40ms, FOV = 240*240mm, 20 slices, voxels 3.75 x 3.75mm, slice thickness = 4.5 mm, and inter-slice interval = 1.0mm. Each subject performed two functional runs, lasting 276 TRs each (552 seconds). During the scan, the subjects performed a syllable identification task, presented in a simple blocked design. In each block, a single syllable was presented once every 2s, for 24s in all. Each stimulus lasted 270ms. At the end of each block, subjects had 5s to press a button, indicating whether they perceived the syllable to be /ra/ or /la/. The auditory stimuli were presented using Avotec MRI-compatible headphones. After each task block, there was a 16s rest block.

Holloway et al. (2010) collected fMRI data from 19 normal adult subjects. Standardised math and reading test scores were available for 16 of those 19 subjects; those are the subjects analysed in the present study. In the nonsymbolic numerical distance-effect task, the same stimuli were used as those reported in two recently published studies (Holloway & Ansari, 2009; Price et al., 2007). Specifically, in this task, participants were asked to determine which of two arrays of white squares contained the larger numerosity. Each nonsymbolic trial matched the number-pair parameters of a corresponding symbolic trial, e.g. a symbolic number comparison of 3 vs. 7 would correspond to a nonsymbolic comparison of 3 squares vs. 7 squares. To control for the possible confound of continuous variables, the density, individual square size, and total area of each array was systematically varied across trials to ensure that numerosity could not be reliably predicted from variables continuous with it. The nonsymbolic control task was created by combining the separate squares into either a shape that resembled a diagonal line or a shape that did not, and the subjects were asked to judge which of two stimuli more closely resembled a diagonal line.

A total of 12 fMRI runs was collected for each participant, three runs for each of the four condi-

tions. Each functional run contained blocks of only one type of comparison task. Functional runs began with 30 s of fixation followed by four 15-second blocks of trials made up of 6 trials each. Each trial was 2.5 s in length. The blocks of trials were separated by 21-second blocks of rest during which subjects were presented with a fixation dot and were not required to make any responses. Functional and structural images were acquired in a 3T Phillips Intera Allegra whole-body MRI scanner (Phillips Medical Systems, The Netherlands) using an 8-Channel Phillips Sense head-coil. A standard BOLD imaging protocol was used to acquire 30 slices in an interleaved order (4 mm thickness, 0.5 mm gap, 80x80 matrix, repetition time (TR): 3000 ms, echo time: 35 ms, flip angle: 90 degrees, field of view 240x240 mm) covering the whole brain. For each functional run, 58 volumes were acquired.

## 2.2 The whole-brain linear regression classifier, and its relation to the Fisher Linear Discriminant and the Pseudo Fisher Linear Discriminant

The linear regression used the voxels' MRI time-courses as input, and the category labels +1 and -1 as the desired outputs. The voxel time courses are made up of individual volumes (i.e. one per TR), not from block-averages. The resulting regression model classifies an input as category +1 or -1 according to whether its output is greater than or less than zero. As described in Raizada et al. (2010), the time-courses were de-trended (i.e. high-pass filtered) and zero-meaned before being entered into the classifier. As described above, the regression weights were calculated using the equation $\mathbf{w} = \mathbf{X}^{\dagger}\mathbf{y}$.

The use of linear regression with the class labels as target outputs is a standard and simple approach for classification, as is its solution using the pseudo-inverse of the data matrix, e.g. pp.184-6 of Bishop (2006). When there are more data points than dimensions, this is equivalent to Fisher's Linear Discriminant (*ibid.* pp.186-190).

The definition of Fisher's Linear Discriminant involves inverting the covariance matrix of the data, which is $p$-by-$p$ where $p$ is the number of dimensions (voxels, in this case). In the context of discriminant analysis, this covariance matrix is often called a scatter matrix. Let $n$ be the number of data points, i.e. the number of brain volumes acquired during the conditions to be compared. The most commonly given form of the equation for linear discriminants involves inverting the within-class scatter matrix, $S_w$. However, the linear regression used here is most directly related to the scatter matrix of *all* of the data, often called the total scatter matrix, $S_t$, or the mixture scatter matrix, $S_m$. The discriminant derived via this total scatter matrix is equivalent to that derived from the within-class scatter matrix, e.g. pp.453-454 of Fukunaga (1990).

When there are more dimensions than data points, i.e. when $p > n$, the total scatter matrix is no longer invertible, as its rank can be no larger than $n$. This is known as the "small sample size" problem (Raudys & Jain, 1991). Many approaches have been suggested for dealing with this

problem. In particular, Duin and colleagues have investigated using the Moore-Penrose pseudo-inverse of the scatter matrix, instead of the regular inverse which no longer exists (Skurichina & Duin, 1999). This approach, which they call the Pseudo Fisher Linear Discriminant, is precisely the one used here.

An interesting and somewhat surprising property of this pseudo-inverse classifier is that it generalises quite well, even when large numbers of irrelevant dimensions are present (Skurichina & Duin, 1999). More recently, the theoretical underpinnings of this method and its relations to a broader class of approaches for dealing with the small sample size problem have been explored by Ye and colleagues (See especially Section 3.4 of Ye et al., 2004).

### 2.3 Cross-validation to prevent overfitting

Typically, when a classifier algorithm is trained on a data-set, the purpose is so that the trained-up classifier can subsequently be used to analyse new data. The trained classifier will perform poorly on the new data if it learned aspects of the training set that are not representative of the classes in the broader population. If that happens, then the classifier will fail to generalise from the training set to subsequent test sets. This is known as over-fitting (Bishop, 2006).

To guard against this, we performed cross-validation. On the /ra/-/la/ data, six-fold cross-validation was used, in the same manner as was carried out in Raizada et al. (2010): the two stimulus conditions (e.g. high-F3 and low-F3) had twelve blocks each, spread across the two runs. For each iteration of the cross-validation, one block from each condition was randomly selected (without replacement) to be used in the test set, resulting in two testing-set blocks corresponding to the two conditions (e.g. F3-high and F3-low). The other ten blocks were used for training. Thus, over the course of the six iterations, every stimulus-block participated in the test set exactly once. Because the blocks were separated from each other by 16 seconds of rest, any potential haemodynamically-induced temporal correlation between the training-set and test-set fMRI data points was reduced. The mean percentage-correct obtained across these six test sets was then calculated, to give the final output.

In the numerical distance effect data, three-fold cross-validation was used, with each of the three functional runs serving in turn as the test set.

The overall cross-validated weight maps for each subject were obtained simply by taking the average of the weight maps from each cross-validation iteration.

## 2.4 A heuristic argument for why a simple linear model may be resistant to degradation by redundant noise dimensions

Some voxels in the brain will contain signal, and many will contain only noise. By "signal", we mean activation which contains information about which stimulus category is being presented, and by "noise" we mean activation which does not contain any such information. In real life we can never know in advance which voxels carry signal and which noise. However, for illustrative purposes let us imagine that we do know this, and let us partition our data matrix, $\mathbf{X}$, into two submatrices $\mathbf{X}_{signal}$ and $\mathbf{X}_{noise}$. By hypothesis, the noise is uncorrelated with the signal, so the expected value of one multiplied by the other is zero $E({\mathbf{X}_{signal}}^T\mathbf{X}_{noise}) = 0$. If the noise has variance $\sigma^2$, then $E({\mathbf{X}_{noise}}^T\mathbf{X}_{noise}) = \sigma^2 I$.

If $\mathbf{X}$ is the data matrix containing $n$ volumes consisting of $p$ voxels each, with each column containing one brain volume stretched out into a $p$-dimensional vector, then the $n$-by-$n$ matrix $\mathbf{X}^T\mathbf{X}$ describes the spatial covariances between the different brain volumes. In the non-degenerate case (e.g. when it is not the case that two volumes are exactly identical), this matrix has rank $n$ and is invertible. Then the pseudo-inverse of $\mathbf{X}$ is $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

The spatial covariance matrix $\mathbf{X}^T\mathbf{X}$ is the crucial term in this formula. Given our hypothetical partitioning $\mathbf{X} = (\mathbf{X}_{signal} \quad \mathbf{X}_{noise})$, we have:

$$\mathbf{X}^T\mathbf{X} = (\mathbf{X}_{signal} \quad \mathbf{X}_{noise})^T(\mathbf{X}_{signal} \quad \mathbf{X}_{noise}) \tag{1}$$

$$= \begin{pmatrix} {\mathbf{X}_{signal}}^T\mathbf{X}_{signal} & {\mathbf{X}_{signal}}^T\mathbf{X}_{noise} \\ {\mathbf{X}_{noise}}^T\mathbf{X}_{signal} & {\mathbf{X}_{noise}}^T\mathbf{X}_{noise} \end{pmatrix} \tag{2}$$

$$E(\mathbf{X}^T\mathbf{X}) = E\begin{pmatrix} {\mathbf{X}_{signal}}^T\mathbf{X}_{signal} & {\mathbf{X}_{signal}}^T\mathbf{X}_{noise} \\ {\mathbf{X}_{noise}}^T\mathbf{X}_{signal} & {\mathbf{X}_{noise}}^T\mathbf{X}_{noise} \end{pmatrix} \tag{3}$$

$$= \begin{pmatrix} E({\mathbf{X}_{signal}}^T\mathbf{X}_{signal}) & 0 \\ 0 & \sigma^2 I \end{pmatrix} \tag{4}$$

$$\tag{5}$$

In other words, in this simple linear system the noise should simply cancel out and should therefore fail to degrade the signal. Clearly this heuristic argument is an over-simplification, which is why the more detailed analysis and simulations presented in Skurichina & Duin (1999) are necessary and important. That paper does indeed suggest that the noise fails to degrade the signal.

# 3 Results

## 3.1 Whole-brain /ra/-/la/ pattern separability correlates with perceptual discriminability

The results of applying the whole brain classifier separability test to the /ra/-/la/ study data above are shown in Figure 1. Despite using the whole brain and plain regression, they are remarkably similar to those previously obtained using a small Heschl's gyrus ROI and an SVM classifier our earlier study (Raizada et al., 2010). The English and Japanese speakers' behavioural abilities to discriminate between differing F2 and F3 stimuli, shown in Fig. 1a, closely mirror the separability of the whole-brain fMRI patterns elicited by those stimuli, shown in Fig.1b. Looking at individual subjects' behavioural and neural measures, in Fig. 1c, it can be seen that this brain-behaviour link also holds across individuals, even after the effect of group membership is partialled out. It is important to note that no information at all about the subjects' behaviour or group membership was given to the classifier: it was provided with only the fMRI data and the condition-labels and onset times of the stimuli that were presented during the scans.

The present brain-behaviour correlation using the whole brain, and the previous result using the Heschl's gyrus ROI, are plotted side-by-side for purposes of comparison in Figure 2.

## 3.2 The pattern-separability difference is specific to /ra/-vs.-/la/

As is shown in Figure 3, the fMRI pattern separability between the speech and baseline conditions does not differ between the English and Japanese subjects. Thus, it is not the case that the Japanese subjects' neural data is somehow intrinsically less separable in general. The separability difference between the groups emerges specifically for the /ra/-/la/ contrast.

## 3.3 Whole-brain distance-effect separability correlates with math scores but not language scores

As is shown in Figure 4, we found a similar result in the distance-effect data: the greater the whole-brain fMRI pattern separability between neural patterns elicited by large-distance pairs and small-distance pairs, the higher were subjects' scores on the arithmetic tests. The correlation with WJ-III (Woodcock-Johnson III) Calculation is significant: $p = 0.0496$, $\rho = 0.498$, two-tailed. The correlation with WIAT (Wechsler Individual Achievement Test) Numerical Operations is marginally significant: $p = 0.0674$, $\rho = 0.468$. The WJ-III Math Fluency correlation is positive, but does not reach significance: $p = 0.144$, $\rho = 0.382$. Scatterplots and regression lines showing this data in more detail can be found in Supplementary Figure S1. The fact that a measure of distance-effect neural activity using the whole-brain correlates with behaviour is noteworthy, as standard fMRI analyses

have found correlated activation exclusively in the parietal cortex.

The observed correlation with behaviour is specific to numerical cognition: the correlations with the reading scores (shown in red) are both weakly negative and non-significant. Thus, it is not the case that greater neural separability on the numerical distance effect task is somehow associated with better behavioural performance across the board.

It should be noted that the scores from different varieties of standardised math tests tend to be highly correlated with each other. The fact that whole-brain fMRI pattern separability correlates to some degree with all three of the math scores should therefore not be taken as meaning that it passes three independent statistical tests. Conversely, although three statistical comparisons are made, a Bonferroni-type multiple-comparisons correction would be inappropriate, again due to the tests' non-independence. If the fMRI pattern separability scores correlate with any one of the math scores, as they indeed do, then they will correlate to some degree with all of them. Although Figure 4 shows three math correlations and two language correlations, it can perhaps best be thought of as simply consisting of two parts: math and language.

The subjects from Holloway et al. (2010) were given two sets of standardised math tests: the Woodcock-Johnson III (WJ-III), and the Wechsler Individual Achievement Test (WIAT). The WJ-III contains two subtests: Calculation, and Math Fluency. The WIAT also contains two: Numerical Operations, and Math Reasoning.

Of these four tests, all but the Math Reasoning test are primarily concerned with basic numerical and arithmetical operations, and would hence be expected to be related to the numerosity processing which is probed in numerical distance effects tasks. In contrast, the Math Reasoning task, as its name suggests, is a test of primarily of logical reasoning, and does not focus on numerosity processing. Consistent with this, two previous studies studying the relation between brain measures and WIAT math tests have found correlations with the Numerical Operations test, but not with Math Reasoning. Isaacs et al. (2001) found this in a voxel-based morphometry study, and van Eimeren et al. (2008) found a similar pattern in a diffusion tensor imaging study of children. Our fMRI data show a similar effect: the correlation between fMRI pattern separability in the non-symbolic numerical distance effect task between small-distance and large-distance number pairs WJ-III Calculation, WIAT Numerical Operations and WJ-III Math Fluency are significant, marginally significant and positive respectively, but there is almost zero correlation with the Math Reasoning scores ($p = 0.96$, $\rho = 0.014$).

### 3.4 Very similar results from using brain-wide logistic regression or linear SVM

One possible concern about the generalisability and robustness of the analyses presented here is the question of whether algorithms more powerful than our very simple linear classifier might produce quite different results. Although the studies of the Pseudo-Fisher Linear Discriminant by

Duin and colleagues (Skurichina & Duin, 1999) show that its simplicity need not prevent it from successfully processing high-dimensional data, stronger evidence would come from carrying out a direct comparison against the results from using more powerful algorithms. We used two such algorithms, each applied to the whole brain at once in exactly the same way as the Pseudo-Fisher Linear Discriminant. The first algorithm is a linear SVM. The LIBLINEAR C-code package (Fan et al., 2008), which comes with a Matlab interface, is able to handle very high dimensional datasets such as a whole-brain analysis on a standard desktop computer. We used the package with its default parameters, including having the penalty parameter at its default value of $C = 1$.

The second is L2-regularised logistic regression. L2-regularisation, unlike L1-regularisation, does not force the classifier weights to be sparse, and therefore produces weight maps which are more directly comparable to those from the linear discriminant. Unlike the linear discriminant, there is no closed-form solution for logistic regression, potentially making its computation numerically difficult in the high-dimensional case. The data forms a large $p$-by-$n$ matrix $\mathbf{X}$, where $p$ is the number of voxels in the whole brain, on the order of tens of thousands, and $n$ is the number of acquired brain volumes, on the order of a few hundred. Hastie & Tibshirani (2004) showed how such cases can exploit a computational shortcut which is frequently used in simpler cases such as calculating the pseudo-inverse: instead of needing to calculate the singular value decomposition of the large $p$-by-$n$ matrix $\mathbf{X}$, we need instead only to deal with the much smaller $n$-by-$n$ matrix $\mathbf{X}^T\mathbf{X}$. This allows logistic regression to be run on the whole brain at once in a fast and memory-efficient manner. We used the Matlab function "svdlr.m" provided as part of the Princeton MVPA toolbox (Detre et al., 2006), which implements the Hastie & Tibshirani (2004) approach.

The results are shown in Supplementary Figure S2. All three algorithms, the very simple Pseudo-Fisher Linear Discriminant and the more complex L2-regularised logistic regression and linear SVM, produced almost identical percentage-correct scores and classifier weight maps. This suggests that the simple parameter-free discriminant, proposed here as a direct non-arbitrary tool for probing brain-behaviour correlations, does indeed seem to be powerful enough to serve its desired purpose, despite its simplicity.

### 3.5 A trade-off between weight-interpretability and comparing the noisiness of different people's representations

A common data pre-processing step in machine learning is to normalise the input features so that they all have variance equal to one (e.g. Aksoy & Haralick, 2001). We explored applying this pre-processing step, and normalised the variance in a single step for each subject, across all volumes in all runs. We found that the normalisation improves the within- and across-subject consistency and interpretability of classifier weight values. However, there is a trade-off, as this variance normalisation also has the unwanted side-effect of slightly reducing the across-subject correlations between classifier-measured neural separability and people's levels of behavioural

performance. The reasons for this are as follows:

For a given weight-size, voxels whose timecourses vary over a wide dynamic range of MRI signal intensity will exert a bigger influence on the overall classification than will voxels whose signals vary over a smaller range, because the large signals from the high-dynamic-range voxels will tend to drown-out the smaller signals from the others. Those smaller-dynamic-range voxels will tend to get drowned out simply because their signal is small, regardless of whether or not that signal would have been able to distinguish between different stimulus conditions. That information within each voxel about the experimental conditions is what we are interested in, but, because of the variability of different voxels' dynamic ranges, the weights assigned by the classifier to the voxels are only loosely related to it. Thus, without normalising the variance of the voxel timecourses, the classifier weights are a poor guide to how much task-relevant information a given voxel contains. Conversely, normalising the timecourses so that they all cover the same dynamic range improves the interpretability of weights, both within and across subjects. This is the positive side of the trade-off.

However, there is also a negative side to the trade-off. When this normalisation is applied to the data, the classifiers' performance improves a little, but the degree to which classifier performance correlates with individual differences in people's behavioural performance goes down slightly. This is because some of the meaningful differences between subjects are removed by the variance-normalisation. Some subjects have brains containing voxels whose dynamic ranges fluctuate wildly: classifier performance for such individuals is affected more by the normalisation than it is for those subjects whose voxels already have similar dynamic ranges. The hypothesis being tested in the present study is that classifier performance will relate to subject's behavioural performance; thus, according to this hypothesis these fluctuating-range subjects would be predicted to perform worse at the behavioural task. However, normalising the voxel dynamic ranges removes this genuine difference between the subjects, and therefore would be predicted to weaken the across-subject correlation between classifier performance and behavioural performance. That is precisely what we in fact observe, as is shown in the Supplementary Material: Supplementary Figs. S3 and S4.

We therefore present the brain-behaviour correlations both without variance-normalisation (i.e. without having removed some of the meaningful inter-subject differences) and also, separately, after variance-normalisation. The post-normalisation correlations are less informative about real brain-behaviour links, and are also slightly weaker, but the classifiers that generated them have the advantage of having more consistent and hence interpretable weights. Using these variance-normalised weight maps, we are able to make some cautious interpretations about the possible roles being played by specific brain areas, and the consistency of those areas across subjects. These are presented in the section immediately below.

We certainly do not consider these analyses to be the last word on the deep problem of how one

person's fine-scale brain activation patterns relate to another's. However, we believe that the interpretations of weight distributions presented here constitute an initial step towards one aspect of that problem.

### 3.6 Across-subject maps of which voxels get assigned the strongest weights by the classifier

The classifier weights from the linear discriminant can be treated as though they were standard GLM contrast images, and entered into a group-level random effects analysis. Thus, the resulting group-level map consists of the t-values of the difference from zero of the average across subjects of the weights in each voxel.

We consider first the data from the numerical distance effect task. When the variance-normalisation described above is applied to the fMRI data before it is entered into the classifier, the group-level weights map does reveal a region whose weights are consistently strong across subjects: the anterior cingulate, as is shown in Fig. 5a. The positive weights correspond in this analysis to voxels which tend to push the classifier towards interpreting an input as having come from a small-distance pair. Such pairs are slightly more difficult for the subject to process, and elicit longer reaction times: the anterior cingulate weights probably reflect this greater mental effort. Such a map should be interpreted with caution: clearly this does not imply that the anterior cingulate is the only region involved in the numerical distance effect task. Standard GLM analyses of the same data confirm the involvement, as expected, of parietal cortex. However, the classifier weights in parietal cortex are too heterogeneous across subjects to be revealed in the analysis presented here.

The negative weights, shown in Fig. 5b, correspond to voxels which tend to push towards classifying an input as having come from a large-distance pair. There is some weak indication of negative weights in the retrosplenial region. The maps are shown thresholded at $p = 0.001$, uncorrected, but should be read as being purely illustrative. Such maps reflect only a small part of what the classifier is doing, and no claims about their statistical significance are being made.

In the /ra/-/la/ task, the interpretation of the weights is slightly more complex. In Raizada et al. (2010), the Heschl's gyrus ROI was found by looking for parts of the brain where the "searchlight" local-neighbourhood fMRI pattern-separability (Kriegeskorte et al., 2006) matched the English vs. Japanese difference in behavioural discriminability. In English speakers, F3 differences are more discriminable than F2 differences, whereas in Japanese speakers both types of formant change are equally poorly discriminable. Thus, the corresponding comparison in terms of searchlight classifier percentage correct scores is to calculate (English F3-minus-F2) minus (Japanese F3-minus-F2). When looking at weight values, instead of percentage-correct values, this formula must be modified slightly. We cannot simply subtract F2 weights from F3 weights: a voxel which has a strongly positive weight when classifying F3 and an equally strong but negative weight when classifying

F2 is equally involved in both formant classifications, but subtracting the negative F2 weight from the positive F3 weight would yield a strongly positive result, misleadingly suggesting that the voxel is much more involved in processing F3 than it is in processing F2. It is therefore necessary to take the absolute values of the weights, before doing the F3-minus-F2 subtraction.

Note that a map of searchlight percentage-correct scores is an entirely different type of measure than a map of brain-wide classifier weights. Because a searchlight sphere centered on one voxel overlaps in large part with the spheres centered on its neighbouring voxels, a map of searchlight percentage-correct values is inherently smooth. This smoothness greatly aids in finding across-subject commonalities, in a manner quite similar to the role played by the explicit spatial smoothing performed in a standard GLM analysis.

For all these reasons, one would expect the group-average map of /ra/-/la/ whole-brain classifier weights to exhibit much less across-subject consistency than the searchlight percentage-correct map calculated in Raizada et al. (2010), and indeed that is the case, as Fig. 6 illustrates. In that figure, both maps are shown at a liberal threshold in order that their broader distributions can be compared, not just the peaks. Thresholding the maps at a low value also helps to highlight the important but often overlooked fact that such peaks usually emerge from large surrounding regions of only slightly subthreshold activation: the appearance of "localised" peaks can obscure the broadly distributed patterns lying just beneath them.

Despite the fact that the classifier-weight map is neither as smooth nor as consistent across subjects as the searchlight percentage-correct map, there are some points of rough similarity them. The whole-brain classifier weight map shows a small peak in the right Heschl's gyrus area, close to the corresponding peak in the searchlight percentage-correct map, and some other similarities can be discerned in the sagittal view, for example in the right middle temporal gyrus.

It is also of interest to look at the positive and negative F3 and F2 weight maps of individual subjects, specifically the "best subject", namely the person whose perceptual ability to hear F3-differences most exceeds their ability to hear F2-differences (this happens to be the English subject En06), and the "worst subject", who shows the opposite behavioural pattern (Japanese subject Jp07). The weight maps for both these subjects are shown in the Supplementary Material, in Supp.Fig. S5. It can be seen that the weights are highly distributed throughout the brain. Although these individual-level maps are interesting to inspect, it is hard to gain much interpretive information from them. For that, the group-level maps shown in Fig. 6 are more appropriate.

## 4 Discussion

When classifying high-dimensional patterns, such as whole-brain images containing tens of thousands of voxels, feature selection has often been found to be necessary. Such methods improve

classifier performance by whittling away the features containing only noise and keeping those which carry signal. In the present case, however, the classifier revealed strong brain-behaviour connections using all the voxels in the brain, without any feature selection at all. Furthermore, the classifier was very simple, and did not use any regularisation or dimension reduction. Despite that, the observed brain-behaviour correlations were strong and specific.

We suggest three factors which may contribute to this: first, the criterion of success here is *not* the more usual measure of how high the classifier's percentage-correct scores are, but instead is how well these classifier scores correlate with the subjects' behaviour. Thus, a classifier which achieves a low percentage-correct may be conveying very accurate information about the underlying neural distributions, if, for example, it is trying to separate the neural patterns elicited by /ra/ and /la/ in the brain of someone who is unable to tell those sounds apart. Two additional factors, described in more detail in the Materials and Methods section above, can be summarised as follows: (i) The simple linear nature of the least squares fit should allow noise-bearing voxels to cancel each other out (ii) The type of linear regression classifier used here has been called the Pseudo Fisher Linear Discriminant in the machine learning literature, and its properties have been studied. In particular, it has been found to perform well when large numbers of redundant feature dimensions are present (Skurichina & Duin, 1999; Ye et al., 2004). This is precisely the situation in a whole-brain analysis.

A very recently published article also linking brain-wide multivoxel patterns to behaviour is that of Marquand et al. (2010), and it provides provides a useful point of comparison for the present study. First, like the work presented here, it shows that the classifier-weight maps which emerge from a brain-wide multivoxel analysis are very highly distributed, and do not exhibit the localised peaks of activation which often emerge from standard univariate analyses. However, the way in which brain-behaviour correlations emerge is different: in Marquand et al. (2010) the subjects' behavioural scores (pain thresholds, in this case) were directly provided to the model, which then explicitly aimed to fit that behavioural data as a function of voxel activations. In contrast, in the present study the classifier algorithm was not provided with any information at all about subjects' behaviour, but instead was given only the neural data and its condition-labels. The separability of the neural data was calculated, yielding just one output value per subject, and without any fitting it emerged that these separability values correlated with subjects' behavioural test scores. Another difference is in the type of statistical model used: Marquand et al. used Gaussian process regression, whereas the present study simply used linear regression.

The overall approach taken by this paper is to try to find as simple and direct a connection between brain and behaviour as possible. Seeking to bridge directly from brain data to behavioural data is certainly not the only possible approach: Kriegeskorte et al. (2008a) has suggested that a potentially powerful way of relating these two domains is by first transforming them all into a common similarity space. This is especially useful when trying to relate radically different types of data, such as human fMRI and monkey neurophysiology (Kriegeskorte et al., 2008b). In the

experimental data considered in the present study, the neural and behavioural data were obtained from the same individual subjects. Given that closer tie, seeking a direct bridge may be warranted.

### 4.1 A cure for "voodoo"?

As has recently been pointed out by Kriegeskorte et al. (2009) and Vul et al. (2009), it is important that statistical tests which are applied to data extracted from an ROI should be distinct from the criteria which were used to define that ROI in the first place. In Raizada et al. (2010), an ROI in right Heschl's gyrus was derived from a groupwise comparison of English versus Japanese speakers. The brain-behaviour correlation which was found to hold in that ROI was across individuals, not across groups, and indeed this correlation remained significant after the effect of group membership was partialled out. Thus, the observed correlation across individuals was distinct from the groupwise contrast that defined the ROI.

The above considerations refer only to the previous study (Raizada et al., 2010). In the present study, there are no ROIs. All analyses are performed across the whole brain at once, without any voxel selection. Thus, the possibility of selection bias does not even arise.

The goal of relating neural activation to behaviour is fundamental to Cognitive Neuroscience, so the concerns raised by Kriegeskorte et al. (2009) and Vul et al. (2009) are in need of a general solution. The surest way to avoid any selection bias is not to do any selection. Previous pattern-based fMRI studies have used a variety of selection and data-reduction steps, such as feature selection of voxels, pre-selecting activated regions, or principal components analysis (e.g., Carlson et al., 2003; Yamashita et al., 2008; Formisano et al., 2008). In a recent overview of the field, Pereira et al. (2009) write the following: "LDA/QDA [Linear Discriminant Analysis / Quadratic Discriminant Analysis] have been used in fMRI before (Carlson et al., 2003; Strother et al., 2002), though always preceded by a dimensionality reduction process (generally a singular value decomposition) in order to train the classifier having fewer dimensions than examples and be able to estimate a covariance matrix." (Note that this quotation is from the Supplementary Information of that paper). The approach presented here avoids the need for any such dimensionality reduction or feature selection process. Moreover, the Pseudo Fisher Linear Discriminant used here successfully handles the case when there are more dimensions than data points; indeed, in both of the data sets presented here the number of dimensions (a whole brain of voxels) exceeds the number of data points (the number of TRs).

## 5 Conclusion

In summary, the results presented above demonstrate that strong brain-behaviour links can be revealed at level of brain-wide multivoxel activation patterns, using plain linear regression. Importantly, this method is free from concerns about non-independent voxel selection or reliance on any specific choice of classifier parameters. In the quest to find connections between brain and behaviour, this offers an approach, unencumbered by arbitrary choices, which is simple and direct.

## Acknowledgements

The authors would like to thank Tanzeem Choudhury, Nikolaus Kriegeskorte and Russ Poldrack for helpful comments on the manuscript.

Kuhl and Raizada were supported by a National Science Foundation (NSF) Science of Learning Center grant (NSF 0354453) to the University of Washington Learning in Informal and Formal Environments (LIFE) Center. Ansari and Holloway were supported by grants from the NSF Science of Learning Center Program (SBE-0354400), the Natural Sciences and Engineering Council of Canada, Canada Foundation for Innovation (CFI), the Ontario Ministry for Research and Innovation (MRI) and the Canada Research Chairs Program.

## References


Aksoy, S. & Haralick, R. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, *22*(5), 563–582.

Anscombe, F. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *J Cogn Neurosci*, *15*(5), 704–17.

Dehaene, S., Molko, N., Cohen, L., & Wilson, A. J. (2004). Arithmetic and the brain. *Current Opinion in Neurobiology*, *14*, 218–24.

Detre, G. J., Polyn, S. M., Moore, C. D., Natu, V. S., Singer, B. S., Cohen, J. D., Haxby, J. V., & Norman, K. A. (2006). The multi-voxel pattern analysis (MVPA) toolbox. *Human Brain Mapping conference presentation.*, *12*. Software available at http://www.csbmb.princeton.edu/mvpa/.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9*, 1871–1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–3.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). Boston: Academic Press.

Hastie, T. & Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, *5*(3), 329–340.

Holloway, I. D. & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: the numerical distance effect and individual differences in children's mathematics achievement. *J Exp Child Psychol*, *103*(1), 17–29.

Holloway, I. D., Price, G. R., & Ansari, D. (2010). Common and segregated neural pathways for the processing of symbolic and nonsymbolic numerical magnitude: an fMRI study. *Neuroimage*, *49*(1), 1006–17.

Isaacs, E. B., Edmonds, C. J., Lucas, A., & Gadian, D. G. (2001). Calculation difficulties in children of very low birthweight: a neural correlate. *Brain*, *124*, 1701–7.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, *103*(10), 3863–8.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–41.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, *12*, 535–540.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., & Mourão-Miranda, J. (2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using gaussian processes. *Neuroimage*, *49*, 2178–2189.

Moyer, R. S. & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–20.

Mussolin, C., Volder, A. D., Grandin, C., Schlögel, X., Nassogne, M., & Noël, M. (2009). Neural correlates of symbolic number comparison in developmental dyscalculia. *Journal of Cognitive Neuroscience*, *10.1162/jocn.2009.21237*.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*, S199–209.

Price, G. R., Holloway, I., Räsänen, P., Vesterinen, M., & Ansari, D. (2007). Impaired parietal magnitude processing in developmental dyscalculia. *Curr Biol*, *17*(24), R1042–3.

Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cereb Cortex*, *20*(1), 1–12.

Raudys, S. & Jain, A. (1991). Small sample-size effects in statistical pattern-recognition - recommendations for practitioners. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, *13*(3), 252–264.

Skurichina, M. & Duin, R. (1999). Regularisation of linear classifiers by adding redundant features. *Pattern Analysis & Applications*, *2*, 44–52.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. *NeuroImage*, *15*(4), 747–71.

van Eimeren, L., Niogi, S. N., McCandliss, B. D., Holloway, I. D., & Ansari, D. (2008). White matter microstructures underlying mathematical abilities in children. *Neuroreport*, *19*, 1117–21.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Yamashita, O., Sato, M.-a., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, *42*(4), 1414–29.

Ye, J., Janardan, R., Park, C. H., & Park, H. (2004). An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, *26*(8), 982–994.

# Figures and captions

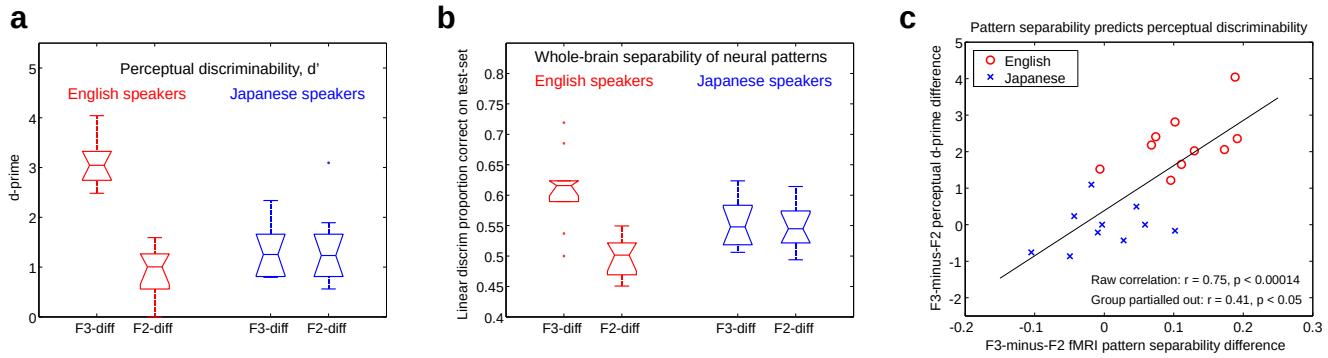


Figure 1: *Relations between fMRI pattern separability and behavioural discriminability in English and Japanese subjects listening to the syllables /ra/ and /la/.* **(a):** *The English and Japanese speakers differ in their behavioural ability to perceive changes in either the F3 formant or the F2 formant. For English speakers, an F3 difference corresponds to a phonetic category change between /ra/ and /la/, and hence is highly discriminable. F2 differences do not produce any category change, and English speakers are correspondingly less able to hear them. For Japanese speakers, neither an F3 nor an F2 difference corresponds to a category change, and hence neither is easily discriminable.* **(b):** *The brain-wide multivoxel fMRI pattern separability, as measured using the simple linear regression classifier, shows a pattern that is strikingly similar to people's behavioural discrimination scores.* **(c):** *This correlation between fMRI pattern separability and behavioural discriminability holds not only across groups, but also across individuals.*

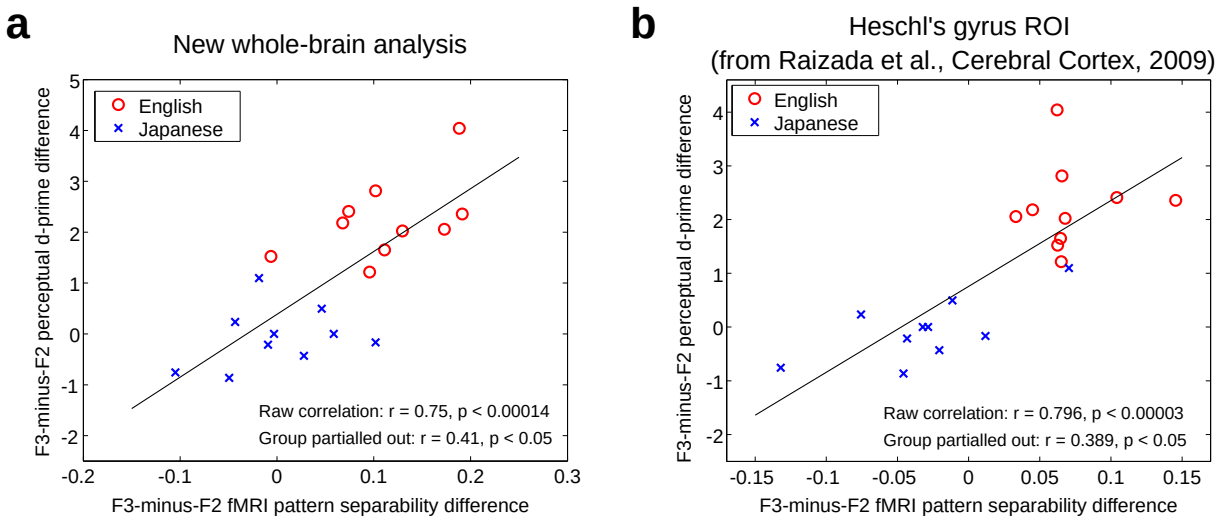


Figure 2: *The brain-behaviour correlation which emerges from the whole-brain classifier analysis presented here* **(a)** *is remarkably similar to that obtained in Raizada et al. (2010), in which the voxels were restricted to a Heschl's gyrus ROI* **(b)**. *Indeed, the correlation after group-membership has been partialled out is actually slightly stronger in the whole-brain case. This shows that a whole-brain analysis can extract behaviourally relevant information despite the presence of large numbers of uninformative voxels.*

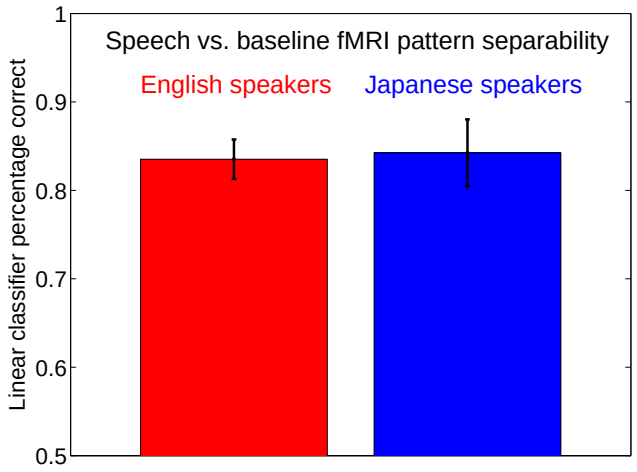


Figure 3: *A control: the fMRI pattern separability between the speech and baseline conditions does not differ between the English and Japanese subjects. Thus, it is not the case that the Japanese subjects' neural data is somehow intrinsically less separable in general. The separability difference between the groups emerges specifically for the /ra/-/la/ contrast. The heights of the bars show the means, and the error bars show the standard errors of the mean.*

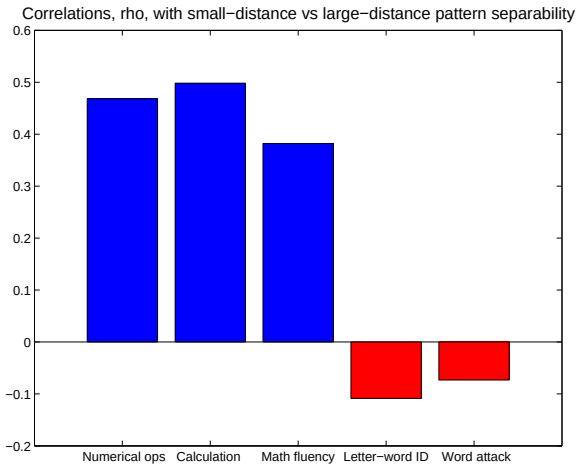


Figure 4: *Correlations between fMRI pattern separability in a non-symbolic numerical distance effect task between small-distance and large-distance number pairs and standardised test scores on mathematics and reading tasks. The correlation with WJ-III (Woodcock-Johnson III) Calculation is significant: $p = 0.0496$, $\rho = 0.498$, two-tailed. The correlation with WIAT (Wechsler Individual Achievement Test) Numerical Operations is marginally significant: $p = 0.0674$, $\rho = 0.468$. The WJ-III Math Fluency correlation is positive, but does not reach significance: $p = 0.144$, $\rho = 0.382$. Scatterplots and regression lines showing this data in more detail can be found in Supplementary S1. The correlations with the reading scores (shown in red) are both weakly negative and non-significant.*

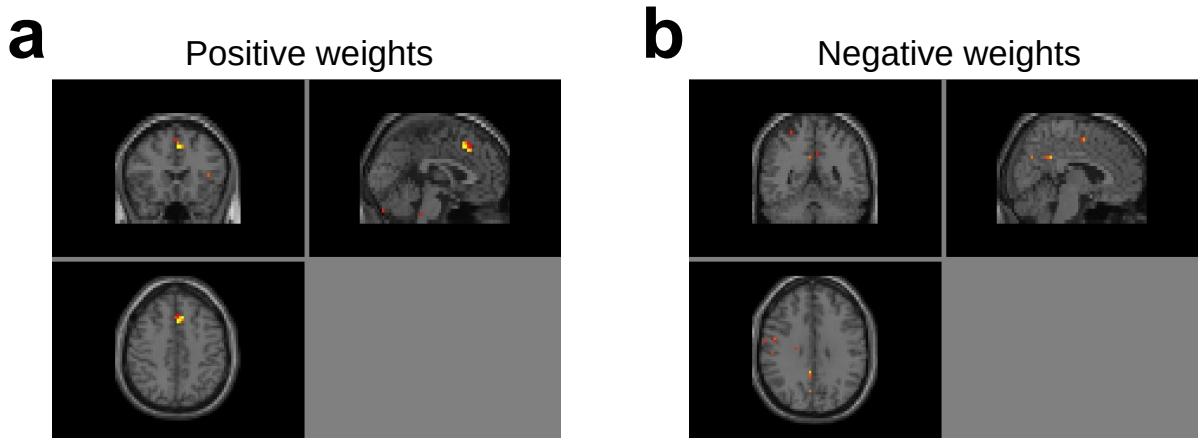


Figure 5: *The group-level random effects map of the classifier weights from the numerical distance-effect task does reveal a region whose weights are consistently strong across subjects: the anterior cingulate, as is shown in* **(a)**. *The positive weights correspond in this analysis to voxels which tend to push the classifier towards interpreting an input as having come from a small-distance pair. The negative weights, shown in* **(b)**, *correspond to voxels which tend to push towards classifying an input as having come from a large-distance pair. There is some weak indication of negative weights in the retrosplenial region. Maps are shown thresholded at p = 0.001, uncorrected, but should be read as being purely illustrative. Such maps reflect only a small part of what the classifier is doing, and no claims of statistical significance are being made. Note that these maps are derived from applying variance-normalisation to the data before entering it into the classifier, as described in the main text.*

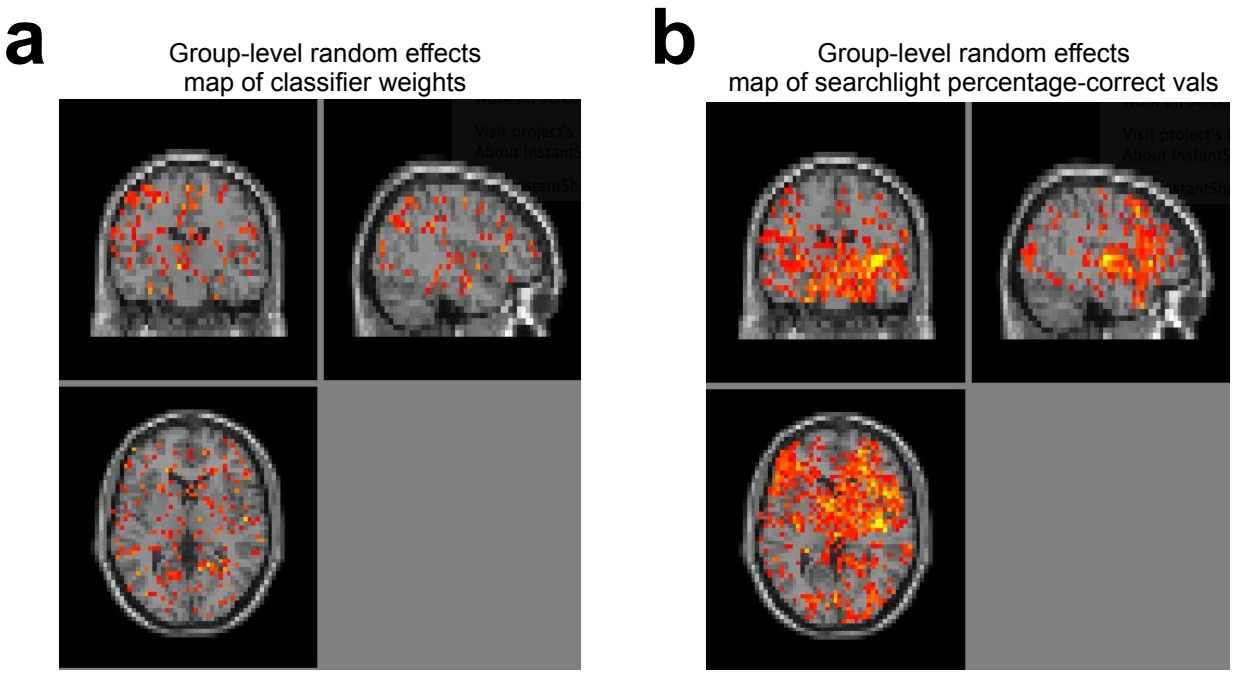


Figure 6: *Comparison between the group-average map of /ra/-/la/ whole-brain classifier weights,* **(a)**, *and the searchlight percentage-correct map calculated in Raizada et al. (2010),* **(b)**. *There are some points of rough similarity between the maps: the whole-brain classifier weight map shows a small peak in the right Heschl's gyrus area, close to the corresponding peak in the searchlight percentage-correct map, and some other similarities can be discerned in the sagittal view, for example in the right middle temporal gyrus. Both maps are shown at a liberal threshold (T=1) in order that their broader distributions can be compared, not just the peaks. Thus, the maps should be viewed as only having illustrative value: no claims of statistical significance are being made.*