NeuroImage 49 (2010) 603–611

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Dynamic Causal Modeling applied to fMRI data shows high reliability

Brianna Schuyler [a,b], John M. Ollinger [a], Terrence R. Oakes [a,d], Tom Johnstone [a,e], Richard J. Davidson [a,c,*]

[a] Waisman Laboratory for Brain Imaging and Behavior, University of Wisconsin-Madison, USA
[b] Neuroscience Training Program, University of Wisconsin-Madison, USA
[c] Psychology Department, University of Wisconsin-Madison, USA
[d] TBI National Capital Neuroimaging Consortium, Washington, DC, USA
[e] Centre for Integrative Neuroscience & Neurodynamics, School of Psychology and CLS, University of Reading, UK

ARTICLE INFO

Article history:
Received 11 May 2009
Revised 2 July 2009
Accepted 9 July 2009
Available online 18 July 2009

Keywords:
Dynamic Causal Modeling
FMRI
Test–retest
Reproducibility
Reliability

ABSTRACT

Sensitivity, specificity, and reproducibility are vital to interpret neuroscientific results from functional magnetic resonance imaging (fMRI) experiments. Here we examine the scan–rescan reliability of the percent signal change (PSC) and parameters estimated using Dynamic Causal Modeling (DCM) in scans taken in the same scan session, less than 5 min apart. We find fair to good reliability of PSC in regions that are involved with the task, and fair to excellent reliability with DCM. Also, the DCM analysis uncovers group differences that were not present in the analysis of PSC, which implies that DCM may be more sensitive to the nuances of signal changes in fMRI data.

Introduction

As the analysis tools of functional MRI data evolve into methods of greater complexity, ongoing critical examination of their sensitivity, specificity, and reproducibility is vital to ensure that interpretations of the data do not exceed the limitations of the data itself. In the past few years, several limitations have been examined. The accuracy of transforming data to a common space has been called into question because of the assumptions made about uniform brain anatomy (Devlin and Poldrack, 2007; Feredoes and Postle, 2007; Goldman-Rakic, 2000; Saxe et al., 2006). In addition, models of the physiological basis of the BOLD signal and its neurological underpinnings have been under continual examination (Aguirre et al., 1998; Handwerker et al., 2004; Leontiev and Buxton, 2007; Raichle and Mintun, 2006). And finally, leading researchers (Logothetis, 2008) have shown that BOLD activity could be due to inhibitory or excitatory effects.

The need for this reevaluation extends to complex mathematical analysis methods. Emerging data analysis techniques that draw from unrelated fields such as engineering and economics may be conceptually and statistically sound, but their power can only be realized if their assumptions are satisfied, and this is often difficult to do in the context of brain imaging. They must therefore be shown to be reliable and valid when applied to a variety of well-characterized datasets. This is a prerequisite for rigorous interpretation of findings about group differences in brain activation, correlation of brain activation with tasks, subject traits, or subject behavior, or more complex understanding of connectivity of different regions in a single brain. Moreover, reliability and validity are of paramount importance when fMRI is applied to clinical research, clinical diagnosis, and the study of neural changes arising from therapeutic interventions.

Many studies (Aron et al., 2006; Caceres et al., 2009; Cohen and Dubois, 1999; Friedman et al., 2006, 2007; Johnstone et al., 2005; Le and Hu, 1997; Loubinoux et al., 2001; Waites et al., 2005; Wei et al., 2004) have focused on the reliability of fMRI signal, as measured by regional percent signal change (PSC) during sensorimotor, cognitive and affective tasks. The results have suggested that, while not ideal, generally there is evidence for reliability of PSC within a subject across time. These studies calculate reliability using Type 1 Intraclass Correlation Coefficient, the ICC (1,1), based on the work of Shrout and Fleiss (1979). ICC (1,1) is calculated using

$$\mathrm{ICC}(1,1) = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (k-1)\mathrm{WMS}}$$

where BMS is the between-subject mean square, WMS is the within-subject mean square, and k is the number of scans on each subject. Cicchetti's guidelines of ICC (Cicchetti, 2001) classify ICC<0.40 as poor, 0.40–0.59 as fair, 0.60–0.74 as good, and 0.75 to 1.00 as excellent. Some studies report reliabilities in the fair to poor range, but many of the studies of reproducibility of PSC would be classified as 'good.' Loubinoux et al. (2001) report highly reliable activation in ROIs in

* Corresponding author. Waisman Laboratory for Brain Imaging & Behavior, 1500 Highland Avenue, Madison, WI 53703, USA.
E-mail address: rjdavids@wisc.edu (R.J. Davidson).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.07.015

response to a motor task repeated 5 h, one month, and two months apart, and Friedman et al. (2007) find good reliability (average ICC = 0.67) in ROIs involved in a block design sensorimotor task with scans one day apart. In a classification learning task, Aron et al. (2006) find good to excellent reliability using voxel-wise ICCs in regions involved with the learning task scanned a few minutes apart. In contrast, Wei et al. (2004) find reliabilities ranging from poor to excellent in different regions in response to an auditory 2-back task, with the Brodmann's Area 40 (BA40) producing the highest reliability across scan sessions at least three weeks apart. The current study replicates the finding that the PSC measurement in an ROI has good reliability in scans that are in the same scanning session (Aron et al., 2006; Caceres et al., 2009; Cohen and Dubois, 1999; Friedman et al., 2006, 2007; Johnstone et al., 2005; Le and Hu, 1997; Loubinoux et al., 2001; Waites et al., 2005; Wei et al., 2004).

The main goal of this work is to study the reliability of Dynamic Causal Modeling (DCM) (Friston et al., 2003) and to see how this method compares to PSC. DCM is an example of a creative application of an engineering technique to the analysis of neuroimaging data. The analysis makes use of time series of activation in a priori regions of interest to estimate changing connectivities between those regions, in addition to direct effects of the stimuli on the regions themselves. It is conceptually appealing and has been used to explore several neural models including language (Mechelli et al., 2005; Bitan et al., 2005; Ethofer et al., 2006; Booth et al., 2007; Sonty et al., 2007; Noppeney et al., 2007), motor activation (Grol et al., 2007; Grefkes et al., 2008), face perception (Fairhall and Ishai, 2006), visuospatial attention (Siman-Tov et al., 2007) and changes across development (Cao et al., 2008; Booth et al., 2008).

A full explanation of the mathematical underpinnings of DCM can be found in Friston et al. (2003), but an overview of the basic structure will be presented here. DCM is an effective connectivity method, as opposed to a functional connectivity method. Functional connectivity analyses examine correlations between regions without regard to causation, while effective connectivity analyses can infer causation within a model. DCM asserts that a change in neuronal firing in one region can be caused by external driving inputs, connections to other regions, and contextual modulation of those inter-regional connections. This is shown explicitly with the formula

$$\dot{z} = \left(\mathbf{A} + \sum \boldsymbol{u}_j \mathbf{B}^j\right)\boldsymbol{z} + \mathbf{C}\boldsymbol{u}$$

where $\boldsymbol{z}$ and $\boldsymbol{u}$ are vectors containing the time series of the neural state of the volumes of interest (VOIs) and stimuli presentations, respectively. **A** is a matrix that describes connectivity between different VOIs when the system is in a steady state, **B** is a matrix that describes how that connectivity changes based on the presentation of each stimulus *j*, and **C** is a matrix that describes the effect of each stimulus on each brain region, comparable to a traditional GLM analysis. The model is then extended from the hypothetical neuronal state to the observed BOLD activity using formulas that calculate blood flow induction, volume changes, and subsequent changes in the fraction of oxy- to deoxy-hemoglobin (Stephan et al., 2007a,b, further discussed in David et al., 2008). These formulas employ biophysical parameters as priors, which are entered into a Bayesian estimation procedure (Friston, 2002). This estimation uses an Expectation-Maximization algorithm to estimate the matrices **A**, **B**, and **C**. The estimated indices of **A**, **B**, and **C** can then be used to explore differences in activation between groups.

In the present experiment, we examined the reliability of results obtained from an analysis of PSC and two DCM analyses, and the ability of each method to discriminate between groups attending to different sensory modalities. We found regions of interest which activated in response to the simultaneous presentation of faces and voices, and studied the behavior in those regions in subjects that were asked to attend to either the face or the voice. We studied the ability of the each analysis method to discriminate between the two groups, and we studied reliability by comparing the analysis results in two scans approximately 2 min apart.

Methods

Participants

Forty-two healthy human subjects were recruited through the local newspaper and chosen by phone screening with an MRI compatibility form and the Edinburgh Handedness Survey. The goal was to obtain a sample representative of the "normal" types of subjects recruited as controls from the population at large. Subjects ranged in age from 18 to 50 years, with number and sex balanced within each decade: 18–29 years, 8 males and 6 females; 30–39 years, 6 males and 6 females; and 40–50 years, 8 males and 6 females. Before participating, each subject gave informed consent. UW-Madison's Human Subjects Committee approved the study paradigm. These data are part of the Wisconsin Neuroimaging Tool Evaluation Resource (WINTER) dataset used to explore fMRI methodology and data analysis issues. Further details on this dataset are outlined in Oakes et al. (2005; Johnstone et al., 2006).

Experimental task

Event-related FMRI was used to examine brain activation in response to angry and happy vocal expressions, while participants concurrently viewed either emotionally congruent or discrepant facial expressions, a task similar to that previously used to examine the crossmodal processing of fear expressions (Dolan et al., 2001). Visual stimuli consisted of 16 greyscale faces displaying expressions of anger and 16 expressions of happiness, half of them females, taken from the Karolinska Directed Emotional Faces set (Lundqvist et al., 1998). Auditory stimuli consisted of short phrases (dates and numbers) lasting on average 1 s, spoken with either angry or happy prosody, taken from the Emotional Prosody Speech and Transcripts dataset (Linguistic Data Consortium, Philadelphia, PA, 2002). 16 angry phrases and 16 happy phrases, half of each spoken by female actors, were used. All phrases were normalized to the same mean signal amplitude.

Stimuli were presented for 1 s with an inter-stimulus interval of 15 s. Each of the four different types of stimulus pairs [angry voice + angry face (AA), angry voice + happy face (AH), happy voice + angry face (HA) and happy voice + happy face (HH)] were presented 20 times each in a pseudo-random order, across two scan runs. Participants performed a two-response (angry or happy) discrimination task. Half the participants were randomly selected to make their decision on the basis of the facial expression (Attentional Focus: 'face' condition), while the other half were instructed to make their decision on the basis of the vocal expression (Attentional Focus: 'voice' condition). Apart from the instruction to attend to either visual or auditory stimuli, all participants performed the identical task, with identical stimulus pairs being presented. Participants were instructed to press one button on a 2-button response pad if the attended stimulus was an angry expression, and to press the other button if the attended stimulus was a happy expression. The matching of buttons to responses was counterbalanced across subjects within each response group. Participants were instructed to respond as quickly and accurately as possible. Stimuli were presented and responses were recorded using EPrime software.

Image acquisition

Images were acquired on a GE Signa 3.0 Tesla MRI scanner device with a quadrature head coil. Anatomical scans consisted of a high resolution 3D T1-weighted inversion recovery fast gradient echo image (inversion time = 600 ms, 256 × 256 in-plane resolution, 240 mm FOV, 124 × 1.2 mm axial slices), and a T2-weighted spin echo image with the same slice position and orientation as the functional images (256 × 256

Table 1
Size and location in MNI space for the four regions studied.

Region	Extent of activation (mm^3)	MNI coordinates		
		x	*y*	*z*
A1	1089	−52	−18	2
V1	931	4	−86	−4
Fusiform	137	36	−52	−22
M1	363	−42	−22	60

Regions are auditory cortex (A1), visual cortex (V1), fusiform gyrus, and motor cortex (M1).

in-plane resolution, 240 mm FOV, 30×4 mm sagittal slices with a 1 mm gap). Two functional scan runs were acquired using a gradient echo EPI sequence (64×64 in-plane resolution, 240 mm FOV, TR/TE/Flip = 2000 ms/30 ms/90°, 30×4 mm interleaved sagittal slices with a 1 mm interslice gap; 290 3D volumes per run).

Analysis

Individual subject data were slice-time corrected in AFNI (Cox, 1996) to correct for temporal offsets in the acquisition of slices (which were acquired in an interleaved slice order). Image distortion was corrected using measured B0 field-maps to shift image pixels along the phase encoding direction in the spatial domain (Jezzard and Balaban, 1995). The data were then transformed to ANALYZE format and the rest of the preprocessing and analysis was done using SPM5. The slices were realigned to the first scan of each session using a 5 mm smoothing kernel and resliced using a 4th degree B-spline interpolation. The functional data from individual subjects were analyzed using a General Linear Model (GLM) with stimulus presentation modeled with a canonical hemodynamic response function, as defined in SPM5. The stimuli were not subdivided on the basis of their emotional valence because the models we were studying deal with strictly sensory regions.

The output of the individual GLM analyses was then transformed to the MNI152 template using FSL's FLIRT (Jenkinson et al., 2002) and volumes of interest (VOIs) were determined from a second level group analysis. In the second level analysis we used the first scan run of each subject and found the clusters that survived an uncorrected voxel-wise threshold of $t>10.13$ ($p<10^{-12}$). The clusters were then transformed back from the MNI-152 template to each subjects' native space using the inverse transformation as calculated in FLIRT. The regions were the left auditory cortex (A1), primary visual cortex (V1), right fusiform gyrus (including the fusiform face area), and a cluster that spans the left primary motor area and (M1) and a portion of the left somatosensory cortex (BA4 and BA6). Details of cluster size and location are given in Table 1, and regions are shown in Fig. 6. These clusters were then entered as VOIs in SPM5 and the first eigenvariates of the timeseries were extracted and employed as the timeseries of the cluster. VOIs from each session were then analyzed in DCM with two different simple models — an auditory model (Fig. 1) and a visual model (Fig. 2). It was hypothesized that activity in the sensory regions would relate to activity in the motor area because participants were instructed to press a button in response to the presentation of the stimulus. We also added stimulus-modulated connections, as it was hypothesized that the connectivity between regions in the model would increase as a result of presentation of relevant stimuli. Both models were applied to subjects that attended to faces and subjects that attended to voices, and the values of the resulting parameter estimates of the models were compared across groups. Reliability of output of the GLM and DCM was measured in SPSS using a single measure Type 1 Intraclass Correlation Coefficient.

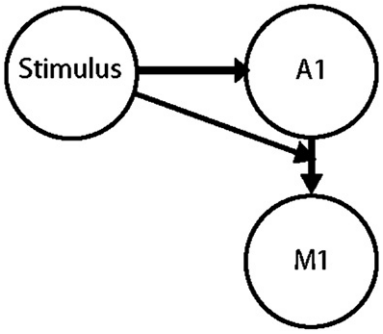

Fig. 1. Schematic illustrating the connections in the auditory model that were tested with DCM. The model posits that the stimulus (simultaneous presentation of face and voice) has a direct effect on the auditory cortex (A1), and a modulatory effect on the connection from the auditory cortex to the motor cortex (M1).

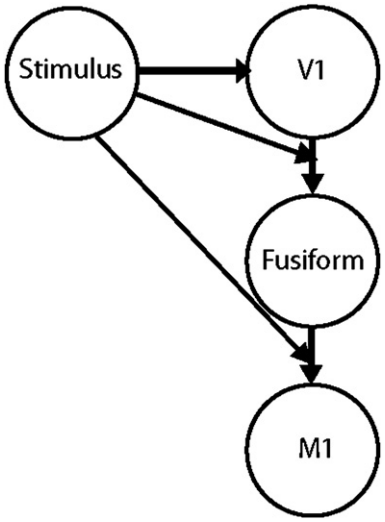

Fig. 2. Schematic illustrating the connections in the visual model that were tested with DCM. The model posits that the stimulus (simultaneous presentation of face and voice) has a direct effect on the visual cortex (V1), and a modulatory effect on the connection from the visual cortex to the fusiform gyrus and the connection from the fusiform gyrus to the motor cortex (M1).

Results

The two separate models, a simple auditory model (Fig. 1) and a more complex visual model (Fig. 2), were entered into the DCM analysis and their parameter estimates were compared with each subject's assigned attentional target (face or voice). In the auditory model it is assumed that the stimulus activates Primary Auditory

Table 2
Reliabilities of model estimates in DCM.

Estimated path coefficient	Scan-rescan reliability (ICC)	
	Visual attentional focus (n = 21)	Auditory attentional focus (n = 20)
Auditory model (Figure 1)		
Stimulus->A1	0.880 (0.909)	0.724 (0.725)
A1->M1	0.616 (0.734)	0.618 (0.647)
Stimulus moderating A1->M1	0.874 (0.878)	0.453 (0.476)
Visual model (Figure 2)		
Stimulus->V1	0.844 (0.865)	0.824 (0.830)
V1->Fusiform	0.825 (0.857)	0.438 (0.471)
Stimulus moderating V1->Fusiform	0.827 (0.915)	0.452 (0.478)
Fusiform->M1	0.918 (0.918)	-0.067 (-0.067)
Stimulus moderating Fusiform->M1	0.638 (0.689)	0.401 (0.485)

Values are color-coded according to Cicchetti's guidelines for reliabilities. ICC<0.40 is poor (black), 0.40–0.59 is fair (red), 0.60–0.74 is good (yellow), and 0.75 to 1.00 is excellent (green). Pearson's r value is shown in parentheses since many parameters increased significantly from Scan 1 to Scan 2.

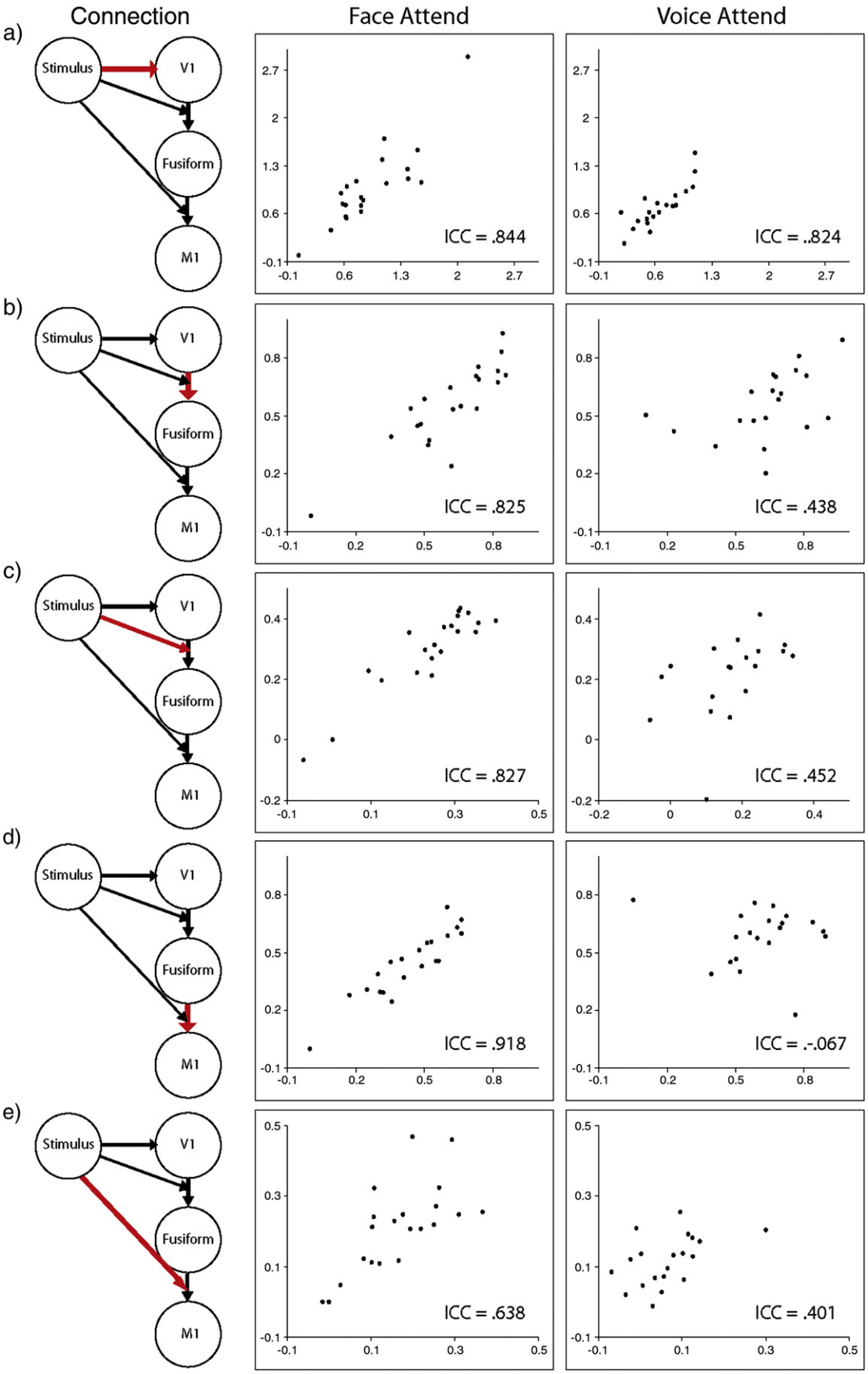

Fig. 3. Visual model DCM values for Scan 1 and Scan 2. Abscissa is Scan 1 and ordinate is Scan 2, for a visual representation of the reliability of the connections in the visual model estimated with DCM. The connections are (a) stimulus effect on V1 (b) V1 connection to fusiform (c) stimulus modulation of V1 connection to fusiform (d) fusiform connection to M1 and (e) stimulus modulation of fusiform connection to M1.

cortex (A1) and activation in A1 then leads to activation in Primary Motor cortex (M1), when the subject indicates his or her discrimination decision with a button press. The visual model follows a similar path, but involves the signal traveling from V1 to the fusiform face area before leading to activation in M1 when the subject executes a button press in response to the emotional face. These two separate models were used to test the reliability of DCM in estimating the parameters of the two different models, and to check if the attended modality of the group could be discriminated by comparing the calculated connectivity values in the two groups. Each subject was assigned only one modality and attended to the same one for both scans.

The parameters calculated in both the auditory and visual DCM models were tested for across-scan reliability in subjects who attended to voice and subjects that attended to face, to test whether the models were more reliable in subjects who were attending to the relevant modality. Reliability results are compiled in Table 2, and plots of the parameter estimates across scans are shown in Figs. 3 and 4. The visual model performed most reliably among subjects who were attending to faces — four of the five parameters had excellent reliability and the last parameter had reliability categorized as "good." The same model, when applied to the subjects that were attending to voices, showed fair to poor reliability in four of the five parameters. Overall, the subjects attending to voice had less reliable activation for both the auditory and visual models, but the disparity is less in the auditory model. The subjects attending to voice had good reliability for the stimulus

Table 3
Reliabilities of percent signal change values.

Region	Visual attentional focus (n = 21)		Auditory attentional focus (n = 20)	
	Intraclass correlation coefficient (ICC)	Mean percent signal change (%)	Intraclass correlation coefficient (ICC)	Mean percent signal change (%)
A1	0.750	0.51±0.20	0.539	0.51±0.19
V1	0.761	0.56±0.31	0.587	0.47±0.28
Fusiform	0.855	0.58±24	0.759	0.41±0.28
M1	0.703	0.24±0.16	0.436	0.31±17

Values are color-coded according to Cicchetti's guidelines for reliabilities. ICC<0.40 is poor (black), 0.40–0.59 is fair (red), 0.60–0.74 is good (yellow), and 0.75 to 1.00 is excellent (green). Percent signal change means and standard deviations are also shown. Regions are auditory cortex (A1), visual cortex (V1), fusiform gyrus, and motor cortex (M1).

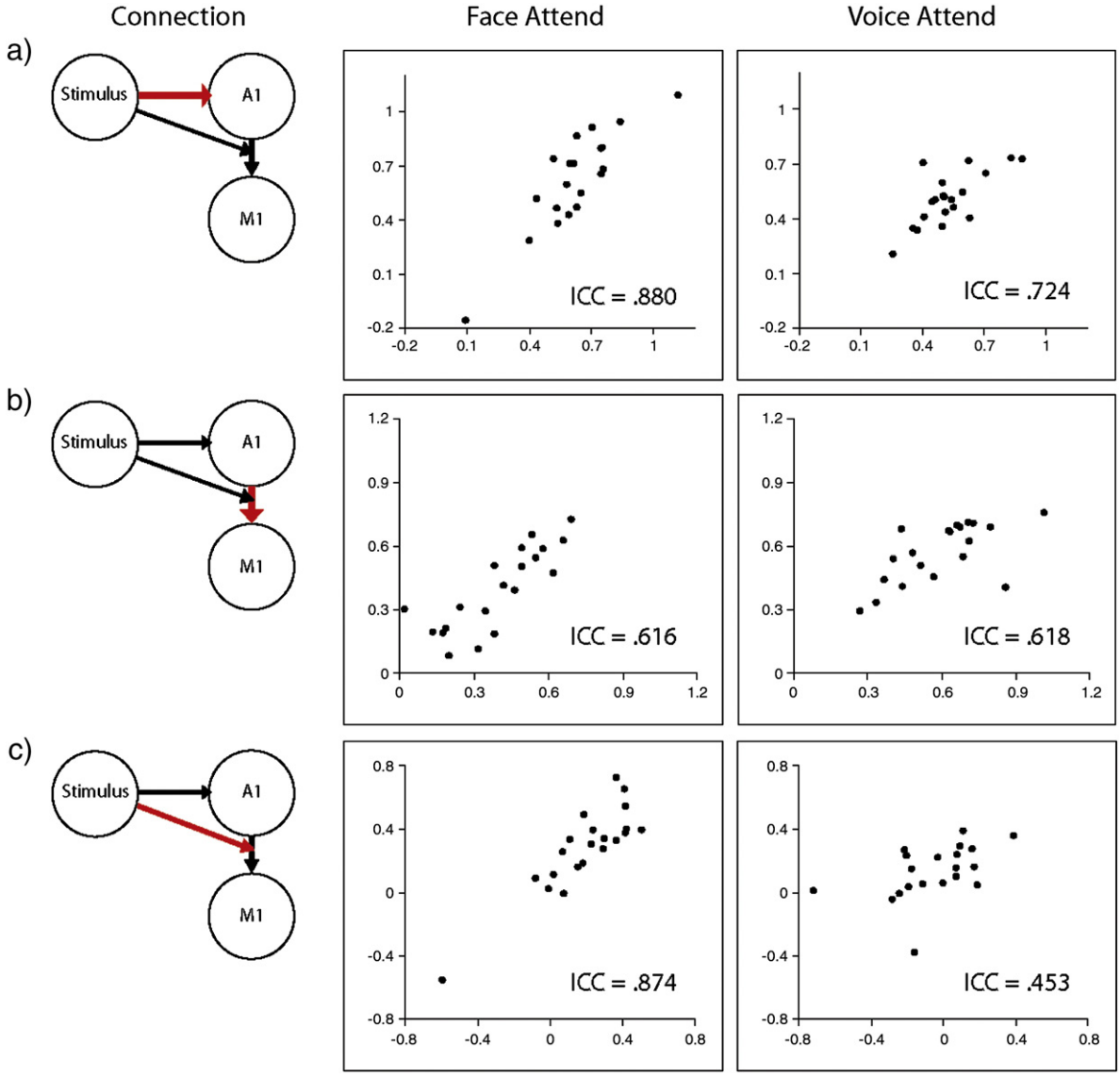

Fig. 4. Auditory model DCM values for Scan 1 and Scan 2. Abscissa is Scan 1 and ordinate is Scan 2, for a visual representation of the reliability of the connections in the auditory model estimated with DCM. The connections are (a) stimulus effect on A1 (b) A1 connection to M1 (c) stimulus modulation of A1 connection to M1.

effect on A1, and the connection from A1 to M1. The stimulus-modulated connection between A1 and M1 had only fair reliability, which is in accordance with the similarly fair reliability of the PSC in those two regions.

To measure the scan–rescan reliability of the PSC we took the clusters that were found in the second level analysis and measured the change in signal during the modeled peak in the hemodynamic response compared to baseline. The measure of reliability, the Intraclass Correlation Coefficient, is a number defined on the interval [−1,1] with reliability improving as the ICC approaches 1. ICC values in the range [−1,0] can effectively be treated as zero reliability. The reliabilities are reported in Table 3, and plots of PSC across scans are shown in Fig. 5. The reliabilities of the PSC for all four ROIs were consistently in the fair to good range for subjects attending to the

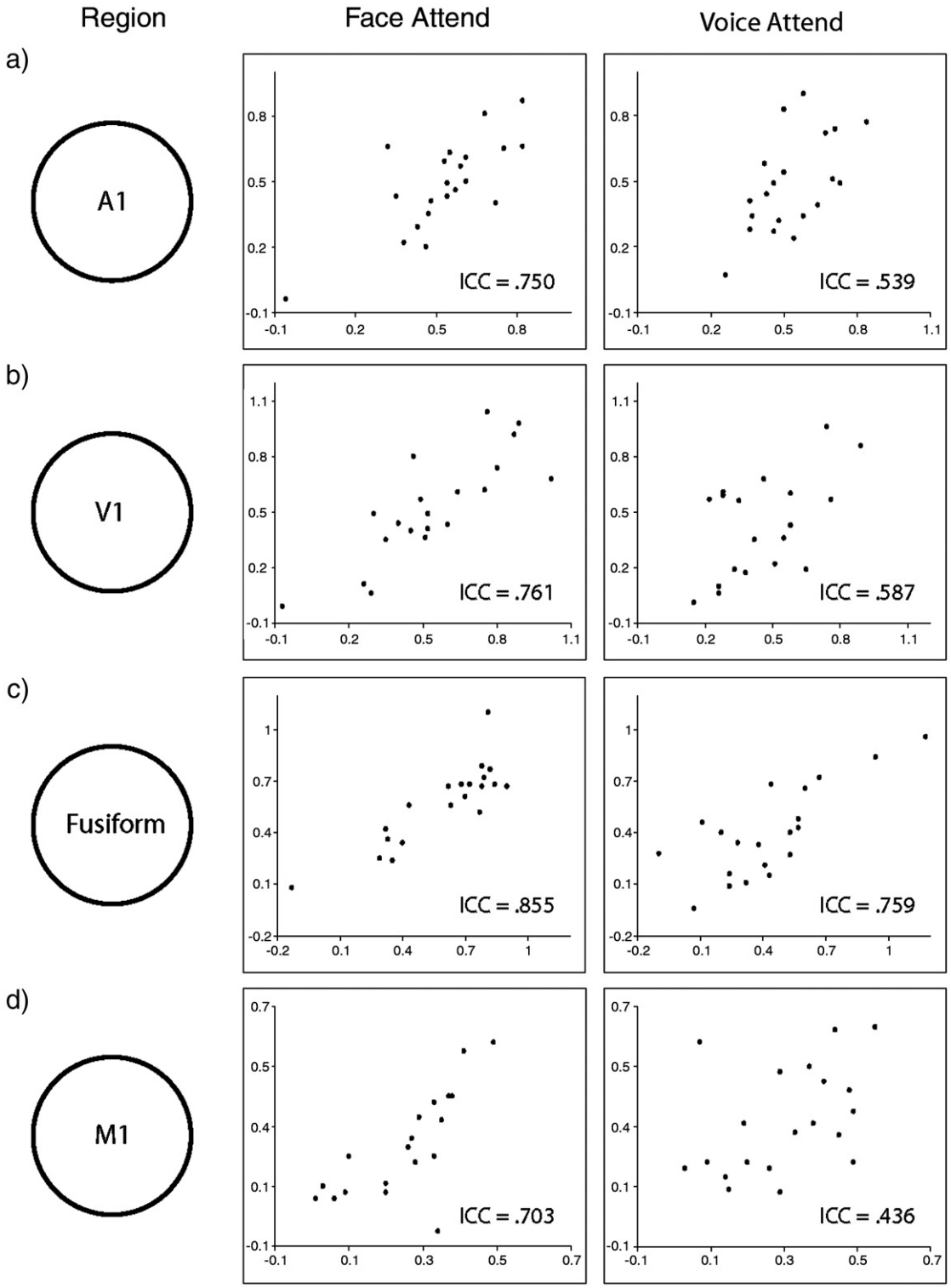

Fig. 5. Percent signal change values for Scan 1 and Scan 2. Abscissa is Scan 1 and ordinate is Scan 2, for a visual representation of the reliability of the percent signal change in the four regions used in the analysis.

voice, and good to excellent range for subjects attending to the face (Fig. 3). It is unclear why A1 and M1 have less reliable PSCs in the subjects attending to the voice than the subjects attending to the face, but this disparity is also present in the DCM analysis.

We then looked for the ability of the DCM analysis to detect differences between groups by testing the interaction of the attentional target with the DCM-estimated connectivity of regions in the two models. In the visual model, three of the five parameters were significantly higher in the subjects that attended to face than the subjects that attended to voice. These were the effect of the stimulus on V1 ($p=0.033$), and both of the stimulus-modulated connections of V1 to fusiform ($p=0.018$) and fusiform to M1 ($p=0.001$). The connection from fusiform to M1 in the absence of the stimulus was significantly higher in subjects attending to voice than subjects attending to face ($p=0.001$) but the reliability of the voice-attending subjects is very low (ICC $=-0.067$) so results with this parameter are questionable. Both stimulus-modulated connections also showed increases from the first to the second scan ($p=0.002$ for V1 connecting to fusiform during stimulus presentation and $p=0.01$ for fusiform connecting to M1 during stimulus presentation). This accords with the fact that task accuracy significantly increased from the first to the second run in subjects who were attending to the face

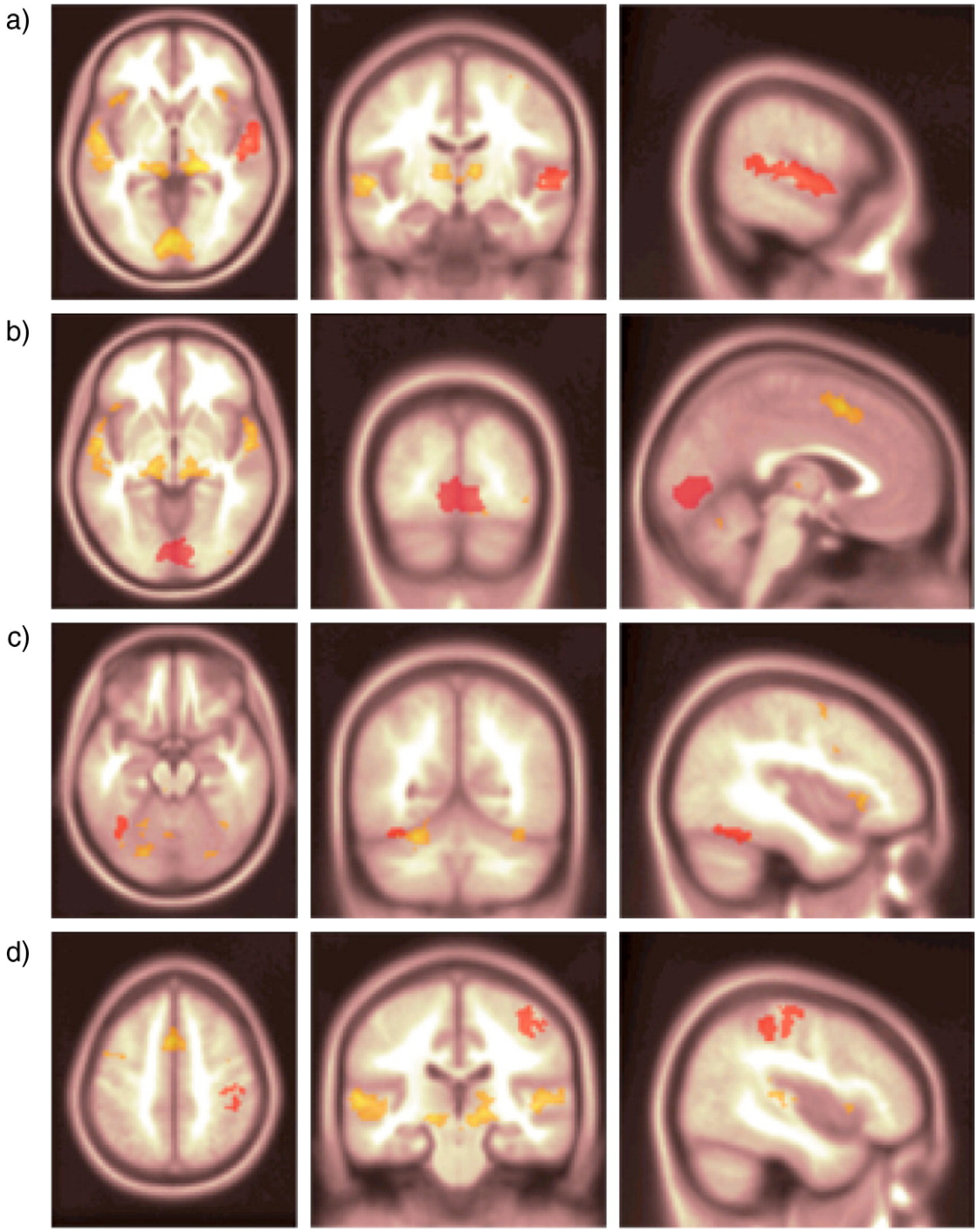

Fig. 6. Statistically defined ROIs in MNI space that were subsequently transformed to the native space of each subject. (a) Left auditory cortex (A1) (b) primary visual cortex (V1) (c) right fusiform gyrus (d) left motor cortex (M1).

(from 77.8% to 86.9%, $p = 0.003$). This ability to measure the connection growing stronger as the task is repeated is something that the PSC approach did not demonstrate in this analysis, and speaks to the ability of DCM analysis to detect subtle changes. The PSC analysis showed no differences over time and significant differences between groups only in fusiform activation ($p = 0.034$).

The auditory model connections calculated in DCM also showed differences between groups. The subjects that attended to the voice showed a stronger connection from A1 to M1 than subjects that attended to face ($p = 0.002$), though subjects that attended to face showed a stronger *stimulus-modulated* connection from A1 to M1 ($p = 0.006$), which is the opposite of what we expected. This might have been caused by the less reliable activity of M1 in the subjects attending to the voice. The stimulus-modulated connection from A1 to M1 held to the same pattern as the stimulus-modulated connections in the visual model, becoming stronger from the first to the second scan ($p = 0.000$). This is again in agreement with the increase in task accuracy across scans. Connectivity values in either model did not predict task accuracy, most likely because accuracy was very high and there was a ceiling effect. Emotion discrimination accuracy was $83.3 \pm 6.7\%$ and $85.5 \pm 8.6\%$ for subjects attending to voice ($n = 21$) and face ($n = 20$), respectively.

Discussion

Many creative methods have emerged to extract information from the massive amounts of data produced by neuroimaging studies in order to characterize the functional role of different brain regions. Both methods evaluated here — one that infers stimulus-related activation by using a General Linear Model to relate brain activity to a predetermined hemodynamic response function (HRF) and another that infers causal relationships between activation in different regions from a mechanistic model, are useful approximations and can produce important insights into neural mechanisms. However, the complexity of the translation from neuronal activity to BOLD signal induction might be better-served by the DCM neuronal to BOLD model that allows for flexibility in modeling the shape of the HRF. Also, the inclusion of connectivity information can produce more complete and sensitive information about differential regional activations and their interactions, which can be more reliable than solely main effects from a GLM. This flexibility in the model, combined with a consideration for regional connectivity and the interaction between regional connectivity and external context, allows a greater ability to detect group differences and nuances of the signal. It can be particularly useful in the analysis of models that are more complex, and regions that are not described well by a canonical HRF. When applied to our data, a simple DCM model performs equally well, if not better, than PSC in reproducibility and ability to distinguish between groups with different attentional targets.

The models we used were relatively simple, dealing with well-characterized regions involved in basic neural functions. An important next step will be to test the reliability of results of analyses performed on regions that are less well characterized, but important to our understanding of more complex psychological processes. It would also be useful to test how well these results reproduce on different days, in different scanners, and with different design paradigms. These questions have been addressed for the reliability of percent signal change (Aron et al., 2006; Cohen and Dubois, 1999; Friedman et al., 2006, 2007; Johnstone et al., 2005; Loubinoux et al., 2001; Waites et al., 2005; Wei et al., 2004) and functional connectivity (Shehzad et al., 2009). It will be valuable to know the power and limitations of other methods as well.

One of the strengths of the current study is the large number of subjects, which allowed us to look more confidently at the reliability and validity of particular methods and models. However, even if practical constraints limit the number of subjects in a particular study, it would be beneficial to do repeated scans of tasks of interest and use the knowledge of reliability when interpreting the results. It is important to know the limitations of current measures and the extent to which a finding is biased by assumed theoretical neural mechanisms and the characteristics of the measurement itself. We have found that, at least when applied to a simple model of well-characterized regions, DCM performs reliably and appears to be sensitive to group effects. Further studies should be done to confirm its ability to handle with less well-defined networks and more complex brain regions. However, it is an excellent example of the potential of methodology adopted from another field to become a powerful and informative method of answering neuroscientific questions through the analysis of neuroimaging data.

Acknowledgments

The authors thank Ron Fisher, Michael Anderle and Kathleen Ores-Walsh for assistance in data collection. This study was supported by NIH grants R01 MH067167, P50-MH084051, and P30-HD03352.

References

Aguirre, G.K., Zarahn, E., D'esposito, M., November 1998. The variability of human, bold hemodynamic responses. Neuroimage 8 (4), 360–369.
Aron, A.R., Gluck, M.A., Poldrack, R.A., February 2006. Long-term test–retest reliability of functional MRI in a classification learning task. NeuroImage 29 (3), 1000–1006.
Bitan, T., Booth, J.R., Choy, J., Burman, D.D., Gitelman, D.R., Mesulam, June 2005. Shifts of effective connectivity within a language network during rhyming and spelling. J. Neurosci. 25 (22), 5397–5403.
Booth, J.R., Wood, L., Lu, D., Houk, J.C., Bitan, T., February 2007. The role of the basal ganglia and cerebellum in language processing. Brain Res. 1133 (1), 136–144.
Booth, J.R., Mehdiratta, N., Burman, D.D., Bitan, T., January 2008. Developmental increases in effective connectivity to brain regions involved in phonological processing during tasks with orthographic demands. Brain Res. 1189, 78–89.
Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A., April 2009. Measuring fMRI reliability with the Intra-Class Correlation Coefficient. NeuroImage 45 (3), 758–768.
Cao, F., Bitan, T., Booth, J., November 2008. Effective brain connectivity in children with reading difficulties during phonological processing. Brain Lang. 107 (2), 91–101.
Cicchetti, D.V., 2001. Methodological commentary the precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsychol. 23 (5), 695–700.
Cohen, M.S., Dubois, R.M., 1999. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J. Magn. Reson. Imaging 10 (1), 33–40.
Cox, R.W., June 1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.
David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., December 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. 6 (12), e315+.
Devlin, J.T., Poldrack, R.A., October 2007. In praise of tedious anatomy. NeuroImage 37 (4), 1033–1041.
Dolan, R.J., Morris, J.S., de Gelder, B., 2001. Crossmodal binding of fear in voice and face. Pro. Natl. Acad. Sci. U. S. A. 98 (17), 10006–10010.
Ethofer, T., Anders, S., Erb, M., Herbert, C., Wiethoff, S., Kissler, J., Grodd, W., Wildgruber, D., April 2006. Cerebral pathways in processing of affective prosody: a Dynamic Causal Modeling study. Neuroimage 30 (2), 580–587.
Fairhall, S.L.L., Ishai, A., December 2006. Effective connectivity within the distributed cortical network for face perception. Cereb. Cortex. 17 (10), 2400–2406.
Feredoes, E., Postle, B.R., April 2007. Localization of load sensitivity of working memory storage: quantitatively and qualitatively discrepant results yielded by single-subject and group-averaged approaches to fMRI group analysis. NeuroImage 35 (2), 881–903.
Friedman, L., Glover, G., Thefbirnconsortiu, M., November 2006. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. NeuroImage 33 (2), 471–481.
Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., Lauriello, J., Lim, K.O., Cannon, T., Greve, D.N., Bockholt, H.J., Belger, A., Mueller, B., Doty, M.J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S.G., 2007. Test–retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29, 958–972.
Friston, K.J., June 2002. Bayesian estimation of dynamical systems: an application to fMRI. Neuroimage 16 (2), 513–530.
Friston, K.J., Harrison, L., Penny, W., August 2003. Dynamic causal modelling. NeuroImage 19 (4), 1273–1302.
Goldman-Rakic, P., May 2000. Localization of function all over again. NeuroImage 11 (5), 451–457.
Grefkes, C., Eickhoff, S.B., Nowak, D.A., Dafotakis, M., Fink, G.R., July 2008. Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. NeuroImage 41 (4), 1382–1394.

Grol, M.J., Majdandzic, J., Stephan, K.E., Verhagen, L., Dijkerman, C.H., Bekkering, H., Verstraten, F.A., Toni, I., October 2007. Parieto-frontal connectivity during visually guided grasping. J. Neurosci. 27 (44), 11877–11887.

Handwerker, D.A., Ollinger, J.M., D'Esposito, M., April 2004. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. NeuroImage 21 (4), 1639–1651.

Jenkinson, M., Bannister, P.R., Brady, J.M., Smith, S.M., 2002. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortion. Magn. Reson. Med. 34 (1), 65–73.

Johnstone, T., Somerville, L., Alexander, A., Oakes, T., Davidson, R., Kalin, N., Whalen, P., May 2005. Stability of amygdala bold response to fearful faces over multiple scan sessions. NeuroImage 25 (4), 1112–1123.

Johnstone, T., van Reekum, C.M., Oakes, T.R., Davidson, R.J., December 2006. The voice of emotion: an fMRI study of neural responses to angry and happy vocal expressions. Soc. Cogn. Affect. Neurosci. 1 (3), 242–249.

Le, T.H., Hu, X., 1997. Methods for assessing accuracy and reliability in functional MRI. NMR in Biomedicine 10 (4–5), 160–164.

Leontiev, O., Buxton, R.B.B., January 2007. Reproducibility of bold, perfusion, and CMRO (2) measurements with calibrated-bold fMRI. Neuroimage 35, 175–184.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453, 869–878.

Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P., Chollet, F, 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. J. Cereb. Blood Flow Metab. 21, 592–607.

Lundqvist, D., Flykt, A., Öhman, A., 1998. The Karolinska Directed Emotional Faces [CD-ROM]. Karolinska Institute, Sweden.

Mechelli, A, Crinion, J.T., Long, S., Friston, K., Lambon Ralph, M.A., Patterson, K., McClelland, J.L., Price, C.J., 2005. Dissociating reading processes on the basis of neuronal interactions. J. Cogn. Neurosci. 17 (11), 1753–1765.

Noppeney, U., Josephs, O., Hocking, J., Price, C.J.J., Friston, K.J.J., July 2007. The effect of prior visual information on recognition of speech and sounds. Cereb. Cortex. 18 (3), 598–609.

Oakes, T.R., Johnstone, T., Ores Walsh, K.S., Greischar, L.L., Alexander, A.L., Fox, A.S., Davidson, R.J., 2005. Comparison of fMRI motion correction software tools. NeuroImage 28, 529–543.

Raichle, M.E., Mintun, M.A., 2006. Brain work and brain imaging. Annu. Rev. Neurosci. 29, 449–476.

Saxe, R., Brett, M., Kanwisher, N., May 2006. Divide and conquer: a defense of functional localizers. Neuroimage 30 (4).

Siman-Tov, T., Mendelsohn, A., Schonberg, T., Avidan, G., Podlipsky, I., Pessoa, L., Gadoth, N., Ungerleider, L.G., Hendler, T., October 2007. Bihemispheric leftward bias in a visuospatial attention-related network. J. Neurosci. 27 (42), 11271–11278.

Shehzad, Z., Kelly, A.M., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Lee, S.H., Margulies, D.S., Roy, A.K., Biswal, B.B., Petkova, E., Castellanos, F.X., Milham, M.P., February 2009. The Resting Brain: Unconstrained yet Reliable. Cereb. Cortex. Advance Access 19 (10), 2209–2229.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Sonty, S.P., Mesulam, X., Weintraub, S., Johnson, N.A., Parrish, T.B., Gitelman, D.R., February 2007. Altered effective connectivity within the language network in primary progressive aphasia. J. Neurosci. 27 (6), 1334–1345.

Stephan, K.E., Harrison, L.M., Kiebel, S.J., David, O., Penny, W.D., Friston, K.J., January 2007a. Dynamic causal models of neural system dynamics: current state and future extensions. J. Biosci. 32 (1), 129–144.

Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., November 2007b. Comparing hemodynamic models with DCM. NeuroImage 38 (3), 387–401.

Waites, A., Shaw, M., Briellmann, R., Labate, A., Abbott, D., Jackson, G., January 2005. How reliable are fMRI–EEG studies of epilepsy? A nonparametric approach to analysis validation and optimization. NeuroImage 24 (1), 192–199.

Wei, X., Yoo, S.S., Dickey, C.C., Zou, K.H., Guttmann, C.R.G., Panych, L.R., March 2004. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. NeuroImage 21 (3), 1000–1008.