Test-Retest Reliability of Effective Connectivity in the Face Perception Network

Stefan Frässle,^{1,2,*} Frieder Michel Paulus,³ Sören Krach,³ and Andreas Jansen^{1,4}

¹Laboratory for Multimodal Neuroimaging (LMN), Department of Psychiatry, University of Marburg, Marburg 35039, Germany

²Department of Child and Adolescent Psychiatry, University of Marburg, Marburg 35039, Germany

³Social Neuroscience Lab | SNL, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck 23538, Germany

⁴Core Facility Brainimaging, University of Marburg, Marburg 35039, Germany

◆ ===== ◆

Abstract: Computational approaches have great potential for moving neuroscience toward mechanistic models of the functional integration among brain regions. Dynamic causal modeling (DCM) offers a promising framework for inferring the effective connectivity among brain regions and thus unraveling the neural mechanisms of both normal cognitive function and psychiatric disorders. While the benefit of such approaches depends heavily on their reliability, systematic analyses of the within-subject stability are rare. Here, we present a thorough investigation of the test-retest reliability of an fMRI paradigm for DCM analysis dedicated to unraveling intra- and interhemispheric integration among the core regions of the face perception network. First, we examined the reliability of face-specific BOLD activity in 25 healthy volunteers, who performed a face perception paradigm in two separate sessions. We found good to excellent reliability of BOLD activity within the DCM-relevant regions. Second, we assessed the stability of effective connectivity among these regions by analyzing the reliability of Bayesian model selection and model parameter estimation in DCM. Reliability was excellent for the negative free energy and good for model parameter estimation, when restricting the analysis to parameters with substantial effect sizes. Third, even when the experiment was shortened, reliability of BOLD activity and DCM results dropped only slightly as a function of the length of the experiment. This suggests that the face perception paradigm presented here provides reliable estimates for both conventional activation and effective connectivity measures. We conclude this paper with an outlook on potential clinical applications of the paradigm for studying psychiatric disorders. *Hum Brain Mapp* 37:730–744, 2016. © 2015 Wiley Periodicals, Inc.

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: Research Foundation of the University of Marburg; Contract grant sponsor: German Research Foundation (DFG); Contract grant numbers: KR3803/2-1, KR3803/7-1; Contract grant sponsor: Else Kröner-Fresenius Stiftung; Contract grant number: 2012_A219; Contract grant sponsor: Förder- und Innovationspool der RHÖN KLINIKUM AG

*Correspondence to: Stefan Frässle, Laboratory for Multimodal Neuroimaging (LMN), Department of Psychiatry, University of

Marburg, Rudolf-Bultmann-Straße 8, 35039 Marburg, Germany. E-mail: fraessle@med.uni-marburg.de

Received for publication 24 August 2015; Revised 26 October 2015; Accepted 10 November 2015.

DOI: 10.1002/hbm.23061

Published online 27 November 2015 in Wiley Online Library (wileyonlinelibrary.com).

Key words: DCM; test-retest reliability; fMRI; face perception; FFA; OFA; computational psychiatry

INTRODUCTION

While early neuroimaging studies have focused on the localization of task-dependent neuronal activation (i.e., functional specialization), it is now generally accepted that any cognitive process rests on the interplay of multiple brain regions (i.e., functional integration). A promising avenue for addressing brain connectivity are computational models that allow examining the neural mechanisms underlying cognitive functions. Over the last decade, such approaches have developed, with dynamic causal modeling (DCM) [Friston et al., 2003] being one of the most frequently used techniques. DCM is a Bayesian framework for investigating the effective connectivity among brain regions based on neuroimaging measures, such as functional magnetic resonance imaging (fMRI). Specifically, DCM makes inference about the directed interactions among brain regions and how these interactions are perturbed by experimental manipulations. Using this technique, researchers have shed light on the neural mechanisms underlying various cognitive processes [e.g., Aukstulewicz and Blankenburg, 2013; Dima et al., 2014; Fairhall and Ishai, 2007; Grefkes et al., 2008; Kellermann et al., 2012; Siman-Tov et al., 2007; Stephan et al., 2007]. Recent efforts have further utilized DCM in clinical settings to gain insights into neurological [e.g., Grefkes et al., 2010; Grefkes and Fink, 2012] and psychiatric disorders [e.g., Almeida et al., 2009; Brodersen et al., 2014; Deserno et al., 2012; Grèzes et al., 2009; Radulescu et al., 2013; Roiser et al., 2013; Schlösser et al., 2008]. In fact, the latter has paved the way for computational psychiatry, which aims to determine aberrant neural mechanisms of psychiatric disorders to (ultimately) guide individualized therapy [Huys et al., 2011; Maia and Frank, 2011; Montague et al., 2012; Stephan et al., 2006; Stephan and Mathys, 2014]. However, despite the potential for understanding normal brain function and the pathophysiology of psychiatric disorders, computational approaches will only prove valuable when experimental paradigms ensure reliable inference on the effective connectivity in individual subjects rather than only at the group level. For instance, if the idiosyncratic properties of such models cannot be determined reliably, they will be of no use for guiding therapy because aberrant pathophysiology—and hence the appropriate treatment—potentially manifests very differently for each individual [Woolrich and Stephan, 2013]. Reliability is thus not only important in cognitive and affective neuroscience, but might become even more important in computational psychiatry. So far, however, systematic analyses on the stability of experimental paradigms for DCM are often lacking and only few studies examined the reliability of DCM for clinical applications.

Here, we address the stability of an fMRI paradigm that has recently been introduced by our group for DCM analyses on the effective connectivity within the face perception network [Frässle et al., 2015a]. Using this paradigm, we investigated the interplay among the two main face-sensitive regions, that is the occipital face area (OFA) [Puce et al., 1996] and the fusiform face area (FFA) [Kanwisher et al., 1997]. Notably, this is the first paradigm to investigate the effective connectivity in the bilateral face perception network, taking into account intra- and interhemispheric interactions. In this framework, we stressed the relevance of interhemispheric integration among the face-sensitive brain regions and provided a mechanistic explanation for the typical right-hemispheric lateralization during face processing. Additional analyses suggested that interhemispheric integration might be related to individual variability in face processing strategies. Hence, this approach might prove valuable for testing the neural mechanisms of face processing at the individual-subject level. Furthermore, as the interhemispheric integration is an important but often neglected aspect of the pathophysiology of psychiatric disorders [e.g., Anderson et al., 2011], the paradigm might also enable deeper insight into psychiatric disorders in which face perception is impaired [e.g., autism spectrum disorders; Vissers et al., 2012].

In the present work, we therefore investigated the test-retest reliability of the face perception paradigm with regard to both conventional activation and effective connectivity measures. Notably, a software comparison on the reliability of DCM has recently demonstrated a marked influence of the prior distributions on the reliability of model parameter estimates [Frässle et al., 2015b]. Specifically, large prior variances afforded less regularization on model inversion and thus led to lower reliability as compared to tighter prior distributions. While several approaches to overcome the influence of the priors on DCM reliability are currently under development (e.g., empirical Bayesian procedures), these are not yet available and, although promising, their utility for improving the stability of model parameter estimation remains to be determined. Based on our previous work, we here addressed this issue in a pragmatic way by relying on DCM8 (as implemented in SPM8, version: 3042) and thus on tighter shrinkage priors. Hence, the reliability estimates in the present study provide an upper bound to the reliability of current DCM versions. In this regard, our findings could serve as a benchmark for future developments aiming to improve the reliability of “state-of-the-art” DCM implementations.

Importantly, for experimental paradigms to be feasible in a clinical setting, not only must the results be reliable, but at the same time, long scanning times need to be

avoided. This is important considering that older, very young, or disabled patients can hardly be expected to lie in the MR scanner for an extensive period of time. Heuristically, this poses a dilemma to MRI paradigms as reliability is typically reduced when less data points (i.e., functional scans) are acquired. In fact, this might be even more severe in the context of DCM, which requires a sufficiently high ratio of data points and model parameters to ensure the stability of the inversion scheme [cf. Daunizeau et al., 2011; Friston et al., 2007; Penny and Roberts, 1999]. This begs an obvious question for the utility of DCM: what is a reasonable trade-off between the reliability of an fMRI paradigm and the urge for short scanning times? This question motivated additional analyses on the reliability of the paradigm as a function of the experiment length.

MATERIALS AND METHODS

Subjects

Twenty-five subjects (11 males, mean age: 23.6 ± 2.8 years) participated in the experiment. All subjects were right-handed [Oldfield, 1971], healthy, and had no history of neurological or psychiatric disease. They performed the identical experiment twice, separated by 53 ± 18 days on average. Prior to the first session, each subject gave informed written consent to participate in the study. The experiment was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee of the Medical Faculty of the University of Marburg.

Experimental Procedure

The experimental paradigm was adapted from a paradigm established for DCM analysis by our group [Frässle et al., 2015a]. Either gray-scale neutral faces (F), objects (O), or scrambled images (S) were presented in the left (LVF) or right visual field (RVF) while subjects fixated a cross in the center of the screen (i.e., 3×2 within-subject experimental design with the factors: stimulus category and hemifield). The face stimuli were taken from the Center for Vital Longevity Face Database [Ebner, 2008]; the object stimuli have been used in previous fMRI studies [e.g., Davies-Thompson and Andrews, 2012]. The scrambled images were the Fourier-randomized versions (i.e., randomizing the phase of the Fourier transform) of the face and object stimuli. Using the Presentation 11.0 software (Neurobehavioral Systems, Albany, CA, <http://www.neurobs.com/>), all stimuli were presented as circular patches (diameter: 4.34 deg) on an MRI-compatible LCD screen (LG SL9000, 60 Hz, 4:3, 1024×786 pix), visible through a mirror attached to the MR head coil.

Stimuli were presented in a block design, with each of the six conditions (i.e., faces, objects, or scrambled images, each in the left or right visual field) being performed seven times during the experiment (42 blocks) in a

pseudo-randomized order. Each stimulus block lasted 14.5 s (i.e., 10 functional whole brain scans) during which 36 stimuli of the same condition were presented (150 ms, ISI = 250 ms). Stimulus blocks were followed by a resting period of the same length where only the fixation cross was shown. Hence, the experiment consisted of 84 blocks in total (i.e., 42 stimulus blocks, 42 fixation blocks). To allow for an investigation of test-retest reliability as a function of the length of the experiment, the experiment was divided into three parts, with subsequent parts being interleaved by 60.9 s (i.e., 42 functional whole brain scans) of rest. The first break was inserted after the 14th fixation block and the second break after the 28th fixation block (see Supporting Information Table S1 for detailed information on how the experimental conditions were split into the three parts). During these breaks, subjects remained inside the scanner and were not allowed to move (as the scanner continued to run). After the last fixation block, a text was shown on the screen for 17.4 s, informing subjects that the end of the experiment has been reached and that they shall not move until the scanner stopped to run. Hence, the total length of the experiment was 22 min 37 s (i.e., 936 functional whole brain scans).

A cover story was used to keep subjects attentive throughout the experiment. Subjects were instructed to judge whether all stimuli within a block were of the correct modality (i.e., faces in face blocks, objects in object blocks, scrambled images in scrambled blocks). They were told that there was a 50% chance that a “wrong” stimulus (e.g., an object presented during a face block) would be presented once during the entire experiment (note, however, that no such “wrong” stimulus was actually shown). After the experiment, subjects reported outside the scanner whether or not they had detected a “wrong” stimulus. None of the subjects indicated the presence of such a “wrong” stimulus for either of the two sessions.

Image Acquisition

Imaging data was acquired at the University of Marburg (Department of Psychiatry) on a 3-Tesla MR scanner (Siemens TIM Trio, Erlangen, Germany). A T_2^* -weighted single-shot gradient-echo echo-planar-imaging sequence (EPI) was used to provide functional images sensitive to the Blood Oxygen Level Dependent (BOLD) contrast (30 slices, Repetition Time (TR) = 1,450 ms, Echo Time (TE) = 25 ms, matrix size 64×64 voxels, voxel size $3 \times 3 \times 4$ mm³, Field of View (FoV) = 192×192 mm², flip angle 90°, descending slice order). In total, 940 functional images were acquired in each session. Additionally, a T1-weighted volume covering the whole brain was obtained during the first session using a magnetization-prepared rapid gradient-echo (3d MP-RAGE) sequence in sagittal plane (176 slices, TR = 1,900 ms, TE = 2.26 ms, matrix size 256×256 voxels, voxel size $1 \times 1 \times 1$ mm³, FoV = 256×256 mm², flip angle 9°).

Functional Imaging Data Analysis

Preprocessing

All analyses of functional imaging data were conducted using SPM8 (version: 3042, Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK; <http://www.fil.ion.ucl.ac.uk>) and Matlab (MathWorks). After discarding the first four images of each session, the following preprocessing steps were conducted for each subject: First, functional images from both sessions were realigned to the individual mean image. Second, the high-resolution anatomical image was coregistered with the mean image. Third, unified segmentation-normalization [Ashburner and Friston, 2005] was applied to the high-resolution anatomical image. The resulting spatial normalization parameters were then used to normalize the functional images into the Montreal Neurological Institute (MNI) standard space. During spatial normalization, functional images were resampled to a voxel size of $2 \times 2 \times 2 \text{ mm}^3$. Fourth, normalized functional images were spatially smoothed with an isotropic 6 mm full width at half maximum Gaussian kernel.

fMRI analysis

The statistical analysis of BOLD activations was conducted using a first-level General Linear Model (GLM) [Friston et al., 1995; Worsley and Friston, 1995]. Each of the six conditions (i.e., faces, objects, or scrambled images, each presented in the LVF or RVF) was modeled as a block regressor and convolved with the canonical hemodynamic response function (HRF). Both sessions were included as two separate sessions in a single model to allow for conjunction analyses, thus resulting in 12 task regressors. Further, 12 realignment parameters (6 for each session) were entered as nuisance regressors in the GLM to account for movement-related artifacts. Additionally, low-frequency noise was suppressed using a high-pass filter (cut-off: 1/128 Hz).

For each subject, six contrast images (3 for each session) were computed. First, voxels responding more strongly to faces as compared to objects and scrambled images, regardless of the visual field, were identified (" $[2 \times F] - [O + S]$ "). Second, voxels responding to stimuli presented in the left visual field, regardless of the stimulus category, were identified ("LVF"). Third, voxels responding to stimuli presented in the right visual field were identified ("RVF"). The individual contrast images of sessions 1 and 2 were then entered into three random effects group-level analyses (i.e., within-subject ANOVAs with the factor "session", which has two levels) to assess BOLD activity for each contrast across both sessions. Anatomical localization and characterization of the activated brain regions were achieved using the Anatomy toolbox extension within SPM8 [Eickhoff et al., 2005].

Additionally, as a measure of the degree of hemispheric lateralization of face-sensitive activation, a lateralization index (LI) was computed [Jansen et al., 2006]. The LI is given by the following expression

$$LI = \frac{\sum \text{activation}_{\text{left}} - \sum \text{activation}_{\text{right}}}{\sum \text{activation}_{\text{left}} + \sum \text{activation}_{\text{right}}} \quad (1)$$

which yields values between -1 and $+1$, with -1 signaling activation purely in the right hemisphere, whereas $+1$ means activation only in the left hemisphere. We here used the bootstrap procedure implemented in the LI toolbox extension [Wilke and Lidzba, 2007] to estimate the LI for each subject individually. This method takes 100 bootstrapped samples (resampling ratio $k = 0.25$) for the ROIs in the left and right hemisphere for 20 equally sized thresholds ranging from 0 to the maximum t -value. This results in 10,000 possible LI combinations, from which only the central 50% are kept to exclude statistical outliers. For each subject, a representative LI is then calculated by weighting these central 50% LIs with their respective threshold. Using this procedure, the LI was computed for face-sensitive activation in the OFA and FFA. Please note that the LI was calculated within the regions of interest (ROIs) of OFA and FFA, which also entered subsequent DCM analyses. Hence, LIs were only assessed for those subjects, for which all DCM-relevant ROIs could be specified (see below).

Time series extraction

Given the group-level results, six ROIs were selected for DCM analyses, in line with our previous approach [Frässle et al., 2015a]. Face-sensitive regions were located in the inferior occipital gyrus (OFA) and in the lateral fusiform gyrus (FFA), each in both hemispheres. Additionally, left and right primary visual cortex (V1) were selected as the input regions for the stimuli shown in the RVF and LVF, respectively. To account for inter-subject variability in the exact location of the six ROIs, the coordinates were determined for each subject individually. Individual coordinates of the ROIs were determined from the first-level conjunction analyses as follows: OFA and FFA were determined from the conjunction of the face-sensitive contrasts ($[2 \times F] - [O + S]_{\text{session1}} \cap [2 \times F] - [O + S]_{\text{session2}}$), right V1 from the conjunction of the left visual field (LVF_session1 $\cap$ LVF_session2), and left V1 from the equivalent right visual field conjunction (RVF_session1 $\cap$ RVF_session2). The individual coordinates of each ROI were determined as the peak activation close to the respective group-level peak coordinate that fulfilled the following anatomical constraints: OFA had to be located in the inferior occipital gyrus, FFA in the fusiform gyrus, and V1 in Brodmann area 17. For each session and subject separately, time series were then extracted from each ROI as the first eigenvariate of all activated voxels ($P < 0.05$, uncorrected)

within a 4 mm sphere centered on the individual coordinates. Time series were mean-centered and corrected for movement-related variance (by specifying an effects-of-interest F -contrast) for each session.

Dynamic Causal Modeling

DCM [Friston et al., 2003] is a Bayesian framework for investigating the effective connectivity within a neural network based on neuroimaging data. Specifically, DCM describes the dynamics of the neuronal states of the network by unveiling the directed interactions among the network regions (endogenous connectivity), and how experimental manipulations perturb these interactions (modulatory influences) and the neuronal state of the regions (driving inputs). The dynamics of the neuronal states in each region are then translated into predictions of the BOLD activity by means of a hemodynamic forward model [Balloon-Windkessel model; Buxton et al., 1998; Mandeville et al., 1999]. Using a Variational Laplace (VL) approach with Gaussian assumptions on the prior and posterior distributions, the posterior densities of the model parameters (i.e., conditional mean and covariance) can be estimated by maximizing the negative free energy [for detailed information on the VL algorithm, see Friston et al., 2007].

Definition of model space

As the exact structure of the network (i.e., the “true” model) is usually unknown *a priori*, DCM analyses require the definition of a sensible model space, which covers well-founded hypotheses on the underlying neural mechanisms. In the present study, sixteen models were constructed, representing different assumptions on the intra- and interhemispheric integration in the face perception network [Fig. 1; for a full motivation of the model space, see Frässle et al., 2015a]. In brief, identical patterns of endogenous connectivity (A-matrix) and driving inputs (C-matrix) were assumed for all models. Driving inputs were set to modulate neuronal activity in the contralateral V1 (e.g., RVF modulated left V1). Forward and backward intrahemispheric endogenous connections were set between V1 and OFA, and between OFA and FFA [Haxby et al., 2000]. Additionally, reciprocal interhemispheric endogenous connections were set among the homotopic face-sensitive regions [Catani and Thiebaut de Schotten, 2008; Park et al., 2008; Van Essen et al., 1982; Zeki, 1970].

Modulatory influences (B-matrix) differed across the sixteen models and varied along two factors: the context of the modulations on (1) the intrahemispheric forward connections, and (2) the interhemispheric connections. Both intra- and interhemispheric connections could be modulated by either (1) the visual field of the stimulus presentation (S), (2) the perception of faces (F), (3) both the visual field and the perception of faces (F+S), or (4) the percep-

tion of faces but only when the faces were presented in the contralateral visual field (F×S).

Each of the experimental inputs was specified as a block regressor. Please note that experimental inputs were not mean-centered—that is, the parameters of the endogenous connections represent the coupling strengths when inputs are zero. Model inversion was then performed using DCM8 as implemented in SPM8 (version: 3042).

Bayesian model selection and Bayesian model averaging

Random effects Bayesian model selection (BMS) [Stephan et al., 2009] was used to rank models according to their negative free energy, which serves (under Gaussian assumptions) as a lower bound on the log model evidence. Individual model parameter estimates were then assessed by means of Bayesian model averaging (BMA) [Penny et al., 2010] across all models within the standard Occam’s window excluding models with a posterior odds ratio below $P < 0.05$ [Madigan and Raftery, 1994]. At the group level, parameter estimates were entered into summary statistics (one-sample t -tests). The significance of each parameter was assessed at a statistical threshold of $P < 0.05$, Bonferroni-corrected for multiple comparisons within each parameter class.

Test-Retest Reliability of BOLD Activity and DCM Results

Various measures have been described for quantifying the test-retest reliability in neuroimaging studies [e.g., Specht et al., 2003]. Here, the procedure for analyzing the test-retest reliability of BOLD activity and DCM results was consistent with that utilized previously [Frässle et al., 2015b]. Reliability was assessed by means of the intra-class correlation coefficient (ICC), specifically the ICC(3,1) type [Shrout and Fleiss, 1979]:

$$\text{ICC}(3,1) = \frac{\sigma_{\text{between}}^2 - \sigma_{\text{within}}^2}{\sigma_{\text{between}}^2 + \sigma_{\text{within}}^2} \quad (2)$$

which provides a meaningful ratio of within-subject variability across sessions (σ_{within}^2) and between-subject variability ($\sigma_{\text{between}}^2$). Notably, there is no consensus regarding a principled threshold that constitutes acceptable reliability in fMRI [Bennett and Miller, 2010]. Nevertheless, qualitative descriptions have been proposed, which help to put the ICC into perspective: Cicchetti [2001] suggested that reliability is classified as “poor” for $\text{ICC} < 0.4$, as “fair” for $0.4 \leq \text{ICC} < 0.6$, as “good” for $0.6 \leq \text{ICC} < 0.75$, and as “excellent” for $\text{ICC} \geq 0.75$.

First, the test-retest reliability of BOLD activity was addressed by calculating the ICC for each voxel in the brain using the ICC toolbox extension within SPM5 [Caceres et al., 2009]. Additionally, the reliability of the BOLD signal relevant for subsequent DCM analyses was

Figure 1.

Different models varying in the modulatory influences on the intra- and interhemispheric connections in the core face perception network [cf. Frässle et al., 2015a]. Forward and backward intra-hemispheric endogenous connections were set between V1 and OFA, and between OFA and FFA. Additionally, reciprocal interhemispheric connections were set between bilateral OFA, and between bilateral FFA, but not between bilateral V1. Arrows indicate the presence of endogenous connections. While endogenous connectivity and driving inputs (not shown here) were the same for all models, modulatory influences differed. Connections were either modulated by (1) the visual field (S), (2) face

perception (F), (3) face perception and visual field (F+S), or (4) face perception, but only when the stimuli were presented in the respective visual field (F×S). These possibilities exist for intra- and interhemispheric connections. Systematically varying all combinations resulted in 16 distinct models. Endogenous connections that were modulated are colored in blue. Naming of the models: type of interhemispheric modulation followed by the type of intra-hemispheric modulation. LVF = left visual field; RVF = right visual field; FP = face perception; L = left hemisphere; R = right hemisphere. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 2.

Face-sensitive BOLD activity during the first session. **A:** Activation pattern shows regions that were more activated during the perception of faces as compared to objects and scrambled images (differential contrast: $[2 \times F] - [O + S]$). The activation pattern is thresholded at a voxel-level threshold of $P < 0.05$ (FWE-corrected). **B:** Reliability map of the voxel-wise ICC(3,1) values

assessed by computing the ICC for the median contrast value within each DCM-relevant ROI (i.e., OFA, FFA, and V1, each in both hemispheres).

Second, the test-retest reliability of DCM was addressed by computing the ICC for the negative free energy and for the model parameter estimates after BMA. Reliability of the negative free energy served as a measure of the stability of random effects BMS. Reliability of the parameter estimates was assessed by calculating an ICC for each parameter separately (i.e., endogenous connections, modulatory influences, and driving inputs). In a second step, the reliability analysis was restricted to the parameter estimates that deviated strongly from their prior expectations, thus providing substantial effect sizes (i.e., posterior probability greater than 0.95).

RESULTS

BOLD Activity During Face Perception and Its Test-Retest Reliability

Using a random effects group analysis (within-subject ANOVA), we found face-sensitive brain activity in a widespread neural network. Activation was primarily located within the inferior occipital gyrus and the fusiform gyrus, referring to OFA and FFA, respectively (Fig. 2A and Table I). Activation of OFA and FFA was observed in both hemispheres, though lateralized to the right hemisphere (mean lateralization index and standard deviation; OFA: $LI = -0.19 \pm 0.35$,

for the differential contrast $[2 \times F] - [O + S]$. The ICC toolbox extension within SPM5 [Caceres et al., 2009] was used to compute the ICC. OFA = occipital face area, FFA = fusiform face area, P = posterior, A = anterior, L = left hemisphere, R = right hemisphere. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$t_{(22)} = -2.58$, $P = 0.02$; FFA: $LI = -0.24 \pm 0.27$, $t_{(22)} = -4.19$, $P < 0.001$). Additional regions showed face-sensitive activations at the group level (e.g., temporal gyrus, amygdala), but were not entered into later DCM analyses [cf. Frässle et al., 2015a]. Across sessions, we found the test-retest reliability of the BOLD activity in the face-sensitive network (i.e., voxels surviving a threshold of $P < 0.001$ uncorrected, for the face-sensitive contrast of the first session, i.e., $[2 \times F] - [O + S]_{\text{session1}}$) to be fair (Fig. 2B; median ICC = 0.52).

Focusing on the reliability of the DCM-relevant BOLD activity, we calculated an ICC of the median contrast value for each of the ROIs entering DCM analyses. We found the test-retest reliability of the median contrast value to be good for the right OFA ($ICC_{\text{med}} = 0.70$) and excellent for all other face-sensitive ROIs (left OFA: $ICC_{\text{med}} = 0.79$; left FFA: $ICC_{\text{med}} = 0.85$; right FFA: $ICC_{\text{med}} = 0.82$). Similarly, reliability of the median contrast value within the input regions of the DCMs (i.e., V1) was excellent in both hemispheres (left V1: $ICC_{\text{med}} = 0.93$; right V1: $ICC_{\text{med}} = 0.83$). This suggests that the test-retest reliability of the BOLD activity within all DCM-relevant ROIs across both sessions was high, and lay well above the values typically reported in fMRI studies [cf. Bennett and Miller, 2010].

Bayesian Model Selection and Its Test-Retest Reliability

Two subjects were excluded from DCM analyses as they did not show consistent face-sensitive activation in all

TABLE I. Shown are brain regions where the conjoint BOLD activity was greater during the perception of neutral faces as compared to the perception of objects and scrambled images ($[2 \times F] - [O + S]_{\text{session1}} \cap [2 \times F] - [O + S]_{\text{session2}}$)

Cortical region	Hemisphere	MNI coordinates			Cluster size	t-score
		x	y	z		
[2×F] – [O+S]_session1 ∩ [2×F] – [O+S]_session2						
Fusiform gyrus (FFA)	R	40	−58	−16	2,204	18.07
Inferior occipital (OFA)	R	48	−78	−4		13.91
Fusiform gyrus	R	40	−40	−18		13.67
Middle temporal	R	48	−62	16	410	9.62
Superior temporal	R	52	−38	12		6.31
Fusiform gyrus (FFA)	L	−38	−64	−16		11.33
Fusiform gyrus	L	−40	−56	−18	300	8.84
Middle temporal	L	−44	−70	18		7.18
Middle occipital	L	−52	−76	14		5.12
Amygdala	L	−22	−6	−12	68	6.01
Cerebellum	R	8	−62	−10	72	5.98
Inferior occipital (OFA)	L	−36	−86	−12	53	5.97
Precentral	R	50	−6	50	449	5.96
Middle frontal	R	34	0	62		4.96
Inferior frontal	R	48	26	16	71	5.40
Suppl. motor area	R	4	10	50	102	5.04
Cerebellum	L	−8	−70	−20	53	4.63

For all regions MNI coordinates, cluster sizes and *t*-scores are reported. Results are thresholded at a voxel-level threshold of $P < 0.001$, uncorrected, cluster size 50 voxels.

ROIs of the network at the statistical threshold of $P < 0.05$, uncorrected (i.e., one subject was lacking face-sensitive activation in the left and right inferior occipital gyrus, corresponding to left and right OFA, respectively; the other subject was lacking face-sensitive activation in the left inferior occipital gyrus and the left fusiform gyrus, corresponding to left OFA and left FFA, respectively). For each of the remaining subjects, models were inverted and then compared using random effects BMS. We found the F+S/F+S model to be the most likely model in both sessions (expected posterior probability (exp_r): 0.55 in session 1 and 0.58 in session 2; exceedance probability (xp): 1.00 in both sessions). The F+S/F+S model suggests that both intra- and interhemispheric connections were modulated by the visual field of the stimulus (S) and the perception of faces (F). For each model, we calculated an ICC for the negative free energy, which showed excellent reliability (ICCs between 0.90 and 0.91). This suggests that BMS was almost perfectly reliable and, in fact, selected the same winning model in both sessions.

Test-Retest Reliability of the Model Parameter Estimates

The model parameter estimates were calculated by means of BMA over all models within the standard Occam's window excluding models with a posterior odds ratio below $P < 0.05$. Varying across subjects, we found

between 1 and 5 models to fall within the Occam's window. Notably, the F+S/F+S model, which was identified as the most likely model using random effects BMS, contributed to the averaged (BMA) parameter estimates for each subject in both sessions. We found the parameter estimates to be highly consistent at the group level across both sessions (Fig. 3A and Supporting Information Table S2). To quantify the within-subject stability of the averaged parameter estimates, we assessed the test-retest reliability by means of the ICC, calculated separately for each endogenous, modulatory, and driving input parameter. In a second step, we restricted our reliability analyses to those parameters with a substantial effect size (i.e., large deviations from their prior mean; see Methods).

ICCs ranged from 0.00 ("poor") to 0.89 ("excellent") and suggested (almost) good reliability of the parameter estimates on average (median ICC = 0.59; Fig. 3B; see Table II for a list of the ICC of each endogenous and modulatory parameter estimate). When analyzing each connectivity matrix separately, we found a median ICC of 0.61 for the endogenous connectivity estimates, 0.58 for the modulatory parameter estimates, and 0.67 for the driving input estimates (LVF: ICC = 0.75; RVF: ICC = 0.60). Please note that for some modulatory parameters, ICC values were equal to zero (Table II, right two columns). This is a consequence of the BMA procedure utilized in the present study. Specifically, models with the modulatory input F×S fell outside the Occam's window of the BMA procedure for some of the subjects in session 1, and did not

Figure 3.

Reliability of DCM parameter estimates. **A:** Group-level results of the modulatory influences on network connectivity (B-matrix) by the perception of faces in session 1 (top) and session 2 (bottom). The group-level results of the parameter estimates are almost identical across sessions. **B:** ICC values for all model parameter estimates, that is for endogenous (blue), modulatory

(light gray), and driving input parameters (orange), as a function of the absolute connection strength. For parameters showing a substantial effect size (i.e., where the posterior differed considerably from the prior) the edge color of the circles is red. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

contribute at all to parameter estimation in session 2 (cf. Supporting Information Table 2). This explains why the ICC values of those parameters that represented modulatory influences on the intra- or interhemispheric connections by F×S were estimated to be zero.

When restricting the analysis to the parameters that showed a substantial effect size (Fig. 3B, circles with a red edge; Table II, bold numbers), we found an elevation of the test-retest reliability (median ICC = 0.68; range: 0.35–0.89). This could be attributed to an increase in reliability of the modulatory parameter estimates (median ICC = 0.70), whereas the reliability of endogenous parameters and driving inputs remained unchanged.

As suggested by one of our reviewers, we repeated the analyses on the test-retest reliability of model parameter estimation for the most likely F+S/F+S model, instead of the averaged (BMA) parameter estimates. For the F+S/F+S model, the reliability of the parameter estimates (restricted to large effect sizes) was highly similar to the results obtained under the BMA procedure and could again be classified as “good” on average (median ICC = 0.63; range: 0.31–0.88, endogenous: 0.62, modulatory: 0.63, driving inputs: 0.70). Although the reliability of the model parameter estimates was slightly decreased for the F+S/F+S model as compared to the BMA procedure, this difference was not statistically significant (Mann–Whitney *U* test: $Z = 0.20$, $P = 0.42$).

Test-Retest Reliability as a Function of the Length of the Experiment

Test-retest reliability of BOLD activity and DCM results was re-analyzed using only the first two parts of the experiment (i.e., “2/3” of the functional images; see Methods), and using only the first part of the experiment (i.e., “1/3” of the functional images).

We found that the reliability of the median contrast value in the DCM-relevant ROIs decreased as a function of the length of the experiment (Fig. 4A). Reliability was only slightly decreased for the left (2/3: ICC_{med} = 0.94; 1/3: ICC_{med} = 0.90) and right V1 (2/3: ICC_{med} = 0.80; 1/3: ICC_{med} = 0.73), as well as for the right OFA (2/3: ICC_{med} = 0.64; 1/3: ICC_{med} = 0.68). On the contrary, test-retest reliability of the left OFA (2/3: ICC_{med} = 0.66; 1/3: ICC_{med} = 0.63), as well as of the left (2/3: ICC_{med} = 0.77; 1/3: ICC_{med} = 0.38) and the right FFA (2/3: ICC_{med} = 0.64; 1/3: ICC_{med} = 0.51) dropped more substantially, although still remaining reasonably high, and above the values reported in most fMRI studies [cf. Bennett and Miller, 2010]. This suggests that BOLD activity within DCM-relevant ROIs was sufficiently stable across sessions even when considering only a small number of data points.

Irrespective of the number of functional images considered for DCM analyses, BMS remained stable across both sessions and identified the F+S/F+S model as the most

TABLE II. Test-retest reliability of the endogenous and the modulatory parameter estimates after Bayesian Model Averaging (BMA)

Connection	intra-class correlation coefficient (ICC)					
	Endogenous	Modulatory				
		LVF	RVF	Faces	Faces (LVF)	Faces (RVF)
FFA_L → FFA_R	0.85	—	0.65	0.82	—	0.00
FFA_L → OFA_L	0.35	—	—	—	—	—
FFA_R → FFA_L	0.81	0.65	—	0.56	0.00	—
FFA_R → OFA_R	0.42	—	—	—	—	—
OFA_L → FFA_L	0.86	—	0.89	0.77	—	—
OFA_L → OFA_R	0.74	—	0.58	0.45	—	0.00
OFA_L → V1_L	0.46	—	—	—	—	—
OFA_R → FFA_R	0.69	0.79	—	0.75	—	—
OFA_R → OFA_L	0.68	0.72	—	0.58	0.00	—
OFA_R → V1_R	0.47	—	—	—	—	—
V1_L → OFA_L	0.54	—	0.70	0.45	—	—
V1_R → OFA_R	0.47	0.48	—	0.59	—	—

ICCs for the parameters that showed a substantial effect size (i.e., deviated strongly from their prior mean, indicated by a posterior probability larger than 0.95) are printed bold.

likely model (2/3: $\text{exp_r} = 0.54$ in session 1 and $\text{exp_r} = 0.57$ in session 2; $\text{xp} = 1.00$ in both sessions; 1/3: $\text{exp_r} = 0.56$ in session 1 and $\text{exp_r} = 0.58$ in session 2; $\text{xp} = 1.00$ in both sessions). Test-retest reliability of BMS, as quantified by the ICC of the negative free energy,

remained virtually unchanged when adapting the length of the experiment and could still be classified as “excellent” (2/3: ICCs between 0.88 and 0.89; 1/3: ICCs between 0.89 and 0.90; see Supporting Information Table S3). Notably, this was not only true for the negative free energy, but also for the values of the individual accuracy and complexity terms of the negative free energy. Reliability of each term remained almost unchanged as a function of the length of the experiment (see Supporting Information Table S4; exemplarily shown for the F+S/F+S model, which was identified as the most likely model in both sessions).

Similarly, reliability of the model parameter estimates dropped only slightly as a function of the length of the experiment (2/3: median ICC = 0.55; 1/3: median ICC = 0.46), being classified as “fair” on average. When analyzing each connectivity matrix separately, we found reliability of the endogenous connectivity estimates (2/3: median ICC = 0.51; 1/3: median ICC = 0.52) and the modulatory parameter estimates (2/3: median ICC = 0.50; 1/3: median ICC = 0.35) to be decreased, whereas reliability of the driving input estimates was slightly increased compared to the full data (2/3: median ICC = 0.75; 1/3: median ICC = 0.72). Restricting the analysis to the parameters with large effect sizes, reliability was increased and could, on average, be classified as “good” when considering 2/3 of the functional images (median ICC = 0.64; range: 0.16–0.90; see Fig. 4B) and “fair” when considering 1/3 of the functional images (median ICC = 0.57; range:

Figure 4.

Reliability of DCM results as a function of the length of the experiment. **A:** ICCs for the median contrast values of the six regions of interest (i.e., OFA, FFA, and VI, each in both hemispheres) as a function of the experiment length. **B:** ICC values for each model parameter estimate, that is for endogenous (light blue), modulatory (light gray), and driving input parameters (light orange), as well as the median ICC value (red) as a function of

the experiment length. To investigate the effect of the number of scans, data analysis was either performed for the full length of the experiment (i.e., 936 functional volumes), two thirds (i.e., 624 functional volumes) or one third of the length of the experiment (i.e., 312 functional volumes). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

−0.16 to 0.92). For 2/3 of the experiment length, the increase in reliability when restricting the analysis to parameters with large effect sizes was due to an increase in the reliability of the modulatory parameter estimates (median ICC = 0.67), whereas the reliability of the endogenous parameters and driving inputs remained unchanged. For 1/3 of the experiment length, reliability increased due to an increase in the reliability of endogenous (median ICC = 0.57) and modulatory parameter estimates (median ICC = 0.51), whereas the reliability of the driving inputs remained unchanged.

Again, we found the same pattern of reliability as a function of the experiment length for the model parameter estimates of the most likely F+S/F+S model, instead of the averaged (BMA) parameter estimates. Reliability dropped only slightly when considering 2/3 (median ICC = 0.65; range: 0.17–0.89, endogenous: 0.51, modulatory: 0.66, driving inputs: 0.76), and 1/3 of the experiment length (median ICC = 0.54; range: −0.14 to 0.92, endogenous: 0.53, modulatory: 0.52, driving inputs: 0.72). In fact, reliability of the model parameter estimates did not significantly differ between the F+S/F+S model and the BMA procedure (2/3: $Z = -0.01$, $P = 0.50$; 1/3: $Z = 0.22$, $P = 0.41$).

In summary, both conventional activation measures (BOLD activity) and DCM results (i.e., BMS and model parameter estimation) were sufficiently stable across sessions even for a short experiment length of less than 8 min [cf. Bennett and Miller, 2010].

DISCUSSION

This study reports the test-retest reliability of an fMRI paradigm for analyzing the effective connectivity among the core regions of the face perception network (i.e., OFA and FFA) by means of DCM. Specifically, this is the first paradigm to address the *bilateral* network by taking into account the intra- and interhemispheric connectivity among the face-sensitive regions. The relevance of interhemispheric integration during face processing has recently been demonstrated in behavioral [Mohr et al., 2002; Schweinberger et al., 2003], functional connectivity [Davies-Thompson and Andrews, 2012], and effective connectivity studies [Frässle et al., 2015a]. Hence, such a paradigm investigating the intra- and interhemispheric integration is essential for understanding the neural mechanisms underlying face processing, and might prove valuable for gaining new insights into the pathophysiology of psychiatric disorders in which face perception is impaired. For this paradigm, we examined the test-retest reliability of conventional activation measures (fMRI) as well as of model selection and model parameter estimation (DCM) across two sessions in 25 healthy volunteers.

Our analyses suggest that conventional activation measures were highly stable within the DCM-relevant ROIs. Similarly, test-retest reliability was excellent for BMS (as

quantified by the ICC for the negative free energy) and, on average, good for model parameter estimation. Reliability for either of the two properties is of relevance as both can be informative when investigating the neural mechanisms underlying normal cognitive functions and the pathophysiology of psychiatric disorders [Seghier et al., 2010]. Notably, to ensure feasibility of the paradigm, we investigated whether DCM results remain stable for shorter experiment lengths. These analyses clearly demonstrated that the reliability of both BMS and parameter estimation remains satisfactory when the length of the experiment is shortened to 8 min. This is of particular relevance for the application of DCM in clinical settings, which require a reasonable trade-off between the reliability of results and the time needed to acquire the data. Note, however, that there is no consensus regarding a principled threshold that constitutes acceptable reliability in fMRI settings [Bennett and Miller, 2010]. Having said this, the ICCs reported for conventional activation and effective connectivity measures in the present study, regardless of the length of the experiment, lie well above the values typically reported in fMRI experiments [cf. Bennett and Miller, 2010]. There are several reasons that might account for the high ICCs reported in the present study. In principle, the paradigm serves as the main determinant for the reliability of fMRI results, whereas other factors such as scanner noise, coregistration errors, or subject motion have less impact [Gorgolewski et al., 2013]. In this regard, the face perception paradigm utilized in the present study was specifically designed to enable high reliabilities by exploiting a robust and powerful imaging sequence providing high signal-to-noise ratios (SNR) in the face-sensitive regions. Furthermore, face processing is likely to elicit similar cognitive processes in both sessions of the experiment with little habituation or learning effects. In this regard, lower reliability reported for other fMRI paradigms (e.g., working memory, spatial attention) might partly represent “true” differences across sessions due to alterations in the exact underlying cognitive processes. Finally, high test-retest reliability of conventional activation and effective connectivity measures in the present study can be attributed to the proper condition of the MR scanner, which was guaranteed by regular maintenance and quality assurance using MRI phantoms. In summary, the paradigm presented here for DCM analysis provides a reliable approach for inferring effective connectivity in the core face perception network from fMRI data.

While test-retest reliability has been examined frequently for conventional activation measures of fMRI [e.g., Aron et al., 2006; Brandt et al., 2013; Fliessbach et al., 2010; Loubinoux et al., 2001; Plichta et al., 2012], only three studies have so far investigated the reliability of DCM [Frässle et al., 2015b; Rowe et al., 2010; Schuyler et al., 2010]. Initially, Schuyler et al. [2010] reported fair to excellent reliability of model parameter estimates. However, their findings lacked generalizability for a number of reasons, such as the use of consecutive within-subject sessions and

small models, as well as the failure to consider BMS. Rowe et al. [2010] found the test-retest reliability to be high for BMS, but low for model parameter estimation. Recently, a more systematic analysis of DCM's test-retest reliability has been demonstrated in the context of a simple motor task [Frässle et al., 2015b]. Comparing two software versions (classical DCM (cDCM) and DCM10 as implemented in SPM5 and SPM8, respectively), the authors showed that the reliability of DCM results depends on the version utilized and can, in fact, be lower for more recent implementations. Specifically, for cDCM, test-retest reliability was high for both BMS and the model parameter estimates. However, reliability of BMS and model parameter estimation was reduced under DCM10. Additional analyses suggested that the drop in reliability of the model parameters (but not of BMS) is attributable to changes in the prior distributions from cDCM to DCM10. Specifically, high reliability of the model parameters was recovered in DCM10 when using cDCM's prior expectations and prior variances for model inversion, whereas the reliability of BMS remained lower. Frässle et al. [2015b] proposed two possible ways in which such changes in the prior distributions could affect reliability of the model parameters. First, the loose priors in DCM10 exert less regularization on the objective function than the tight priors in cDCM, thus leading to a greater likelihood of local extrema in the objective function. Second, the changes in reliability are directly associated with the wide (uninformative) DCM10 priors, which stresses the need for a change in the choice of priors. Currently, new approaches for DCM are under development, which implement global optimization schemes (e.g., Markov Chain Monte Carlo, Gaussian Process Optimization) and Empirical Bayes [Efron and Morris, 1973; Kass and Steffey, 1989] to address local extrema and the choice of priors as limiting factors for test-retest reliability, respectively. In the present study, we used DCM8 as implemented in SPM8 (version: 3042) for model inversion (which is closer to cDCM than DCM10) and hence our findings are—strictly speaking—only applicable to this software version (i.e., set of priors).

Four important aspects need to be emphasized: First, test-retest reliability quantifies the within-subject stability of DCM results across multiple sessions and, although serving as an important prerequisite, does not provide information on the validity of the results. In fact, validity of DCM for fMRI has been investigated intensively in the context of face validity [Friston et al., 2003; Stephan et al., 2008], construct validity in relation to Structural Equation Modeling [Penny et al., 2004] and large-scale neural networks [Lee et al., 2006], cross-validation against other data modalities [Dima et al., 2009, 2010], and predictive validity [Brodersen et al., 2011; David et al., 2008; Reyt et al., 2010].

Second, our analyses suggest excellent reliability of the negative free energy with little variation of the ICC value across different models. This is in line with observations from Frässle et al. [2015b], which demonstrated excellent

reliability of the negative free energy for model inversion in cDCM. Markedly, excellent reliability of the negative free energy was specific to model inversion in cDCM, as reliability dropped to ICC values around 0.6 in DCM10, even when using cDCM's prior expectations and prior variances. Hence, excellent reliability of the negative free energy is not associated with the tight cDCM priors. Note that this is in contrast to our observations for the reliability of model parameter estimation, for which we revealed a crucial role of the prior distributions on the reliability of the parameter estimates. Consequently, the excellent reliability of the negative free energy most probably results from characteristics of the numerical inversion scheme in cDCM/DCM8, such as differences in the update equations of the gradient ascent. For instance, temporal regularization was introduced in DCM10, but was not present in cDCM and DCM8. This is, however, speculative and one would need to test the modifications from DCM8 to DCM10 explicitly in order to delineate the exact cause of the excellent reliability of the negative free energy reported here. Such a detailed analysis is, however, beyond the scope of this paper and thus remains subject to future research.

Third, while the approach presented in this study provides reliable estimates for both conventional activation and effective connectivity measures in the core face perception network, excluding two subjects (corresponding to an attrition rate of 8%) due to a lack of activation in the relevant ROIs depicts an obvious limitation when being interested in effects at the single-subject level. This might become even more severe when aiming to translate such paradigms into a clinical setting since the attrition rate is likely to be higher in (e.g., older or disabled) patients. Hence, thorough analyses on interindividual differences in face-sensitive activation, and why some volunteers do not show consistent activity in the regions of interest, remain subject to future research.

Fourth, in line with our previous approach [Frässle et al., 2015a], here, we restricted the effective connectivity analyses to a face perception network consisting of OFA and FFA in both hemispheres. Other brain regions involved in face processing, such as the posterior superior temporal sulcus (pSTS), amygdala, or insula [Haxby et al., 2000; Ishai, 2008] were not included in the present model. This is consistent with recent findings suggesting that only OFA and FFA are necessary for pure (passive) face detection or identification [Grill-Spector et al., 2004], and that additional regions merely serve additional, face-related functions (e.g., emotion, gaze). Nevertheless, future studies should provide a richer picture of the mechanisms underlying everyday face processing if these regions are also considered. We therefore aim at developing (and testing) more sophisticated experimental paradigms in future studies (e.g., by including emotional vs. neutral faces or by adopting more realistic, dynamic tasks), which will allow us to build models covering a more extended network.

In summary, the fMRI paradigm presented here provides a reliable approach for investigating effective

connectivity in the core face perception network by taking into account the intra- and interhemispheric integration among the core regions (i.e., OFA, FFA). This approach might therefore prove valuable for understanding face processing at the individual-subject level. Furthermore, our results suggest that such an approach could be useful in a clinical setting as well, addressing alterations in the neural mechanisms underlying psychiatric disorders such as major depression [Almeida et al., 2009; Schlösser et al., 2008], autism spectrum disorders [Visser et al., 2012], or schizophrenia [Friston, 1999; Friston and Frith, 1995]. Particularly, in autism spectrum disorders, patients show impairments in basic face perception abilities [for a review, see Weigelt et al., 2012], which are accompanied by atypical brain activation patterns, especially in the FFA [Schultz, 2005]. These alterations in the face perception system have been suggested to contribute to the social deficits from which patients diagnosed with autism suffer. However, our understanding of these alterations in brain function, and how they contribute to the observed impairments in social interactions remains vague. In this regard, it has recently been suggested that the localization of differences in brain activation might not suffice for a thorough understanding of the pathophysiology of autism; yet, impairments need to be formalized from a network perspective [Visser et al., 2012]. Researchers have thus begun to examine alterations of the network structure in terms of brain connectivity. For instance, functional connectivity studies suggested that abnormal connectivity within the limbic system might contribute to the patients' social impairments [Kleinhan et al., 2008; Monk et al., 2010]. However, the interpretation of these results is limited, as functional connectivity is defined by mere statistical dependencies between network nodes [Stephan, 2004]. In contrast, DCM assesses the effective connectivity among the brain regions and yields biophysically interpretable models that can be linked to the individual symptoms of each patient [Stephan and Mathys, 2014]. Researchers have therefore revealed alterations of the effective connectivity in patients diagnosed with autism [Grèzes et al., 2009; Radulescu et al., 2013], raising hope that computational approaches will be of benefit to psychiatry in the future. In this regard, our results demonstrate that the paradigm presented here may offer an innovative and reliable approach in computational psychiatry (in particular when refining the present approach to adopt more realistic, dynamic tasks, and a more extended network), enabling the investigation of intra- and interhemispheric integration in the face perception network and thus providing deeper insight into the pathophysiology of autism and other psychiatric disorders.

REFERENCES

- Almeida JR, Versace A, Mechelli A, Hassel S, Quevedo K, Kupfer DJ, Phillips ML (2009): Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biol Psychiatry* 66:451–459.

- Anderson JS, Druzgal TJ, Froehlich A, DuBray MB, Lange N, Alexander AL, Abildskov T, Nielsen JA, Cariello AN, Cooperrider JR, Bigler ED, Lainhart JE (2011): Decreased interhemispheric functional connectivity in autism. *Cereb Cortex* 21:1134–1146.
- Aron AR, Gluck MA, Poldrack RA (2006): Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage* 29:1000–1006.
- Ashburner J, Friston KJ (2005): Unified segmentation. *Neuroimage* 26:839–851.
- Auksztulewicz R, Blankenburg F (2013): Subjective rating of weak tactile stimuli is parametrically encoded in event-related potentials. *J Neurosci* 33:11878–11887.
- Bennett CM, Miller MB (2010): How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci* 1191:133–155.
- Brandt DJ, Sommer J, Krach S, Bedenbender J, Kircher T, Paulus FM, Jansen A (2013): Test-retest reliability of fMRI brain activity during memory encoding. *Front Psychiatry* 4:163.
- Brodersen K, Schofield T, Leff A, Ong C, Lomakina E, Buhmann J, Stephan K (2011): Generative embedding for model-based classification of fMRI data. *PLoS Comput Biol* 7:e1002079.
- Brodersen KH, Deserno L, Schlagenhaut F, Lin Z, Penny WD, Buhmann JM, Stephan KE (2014): Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin* 4: 98–111.
- Buxton R, Wong E, Frank L (1998): Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magn Reson Med* 39:855–864.
- Caceres A, Hall D, Zelaya F, Williams S, Mehta M (2009): Measuring fMRI reliability with the intra-class correlation coefficient. *Neuroimage* 45:758–768.
- Catani M, Thiebaut de Schotten M (2008): A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex* 44:1105–1132.
- Cicchetti D (2001): The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *J Clin Exp Neuropsychol* 23:695–700.
- Daunizeau J, David O, Stephan K (2011): Dynamic causal modeling: A critical review of the biophysical and statistical foundations. *Neuroimage* 58:312–322.
- David O, Guillemain I, Sallet S, Rey S, Deransart C, Segebarth C, Depaulis A (2008): Identifying neural drivers with functional MRI: An electrophysiological validation. *PLoS Biol* 6:2683–2697.
- Davies-Thompson J, Andrews TJ (2012): Intra- and interhemispheric connectivity between face-selective regions in the human brain. *J Neurophysiol* 108:3087–3095.
- Deserno L, Sterzer P, Wüstenberg T, Heinz A, Schlagenhaut F (2012): Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. *J Neurosci* 32:12–20.
- Dima D, Roiser JP, Dietrich DE, Bonnemann C, Lanfermann H, Emrich HM, Dillo W (2009): Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *Neuroimage* 46:1180–1186.
- Dima D, Dietrich DE, Dillo W, Emrich HM (2010): Impaired top-down processes in schizophrenia: A DCM study of ERPs. *Neuroimage* 52:824–832.

- Dima D, Jogia J, Frangou S (2014): Dynamic causal modeling of load-dependent modulation of effective connectivity within the verbal working memory network. *Hum Brain Mapp* 35:3025–3035.
- Ebner N (2008): Age of face matters: Age-group differences in ratings of young and old faces. *Behav Res Methods* 40:130–136.
- Efron B, Morris C (1973): Stein's estimation rule and its competitors—Empirical Bayes approach. *J Am Stat Assoc* 68:117–130.
- Eickhoff S, Stephan K, Mohlberg H, Grefkes C, Fink G, Amunts K, Zilles K (2005): A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* 25:1325–1335.
- Fairhall SL, Ishai A (2007): Effective connectivity within the distributed cortical network for face perception. *Cereb Cortex* 17:2400–2406.
- Fliessbach K, Rohe T, Linder N, Trautner P, Elger C, Weber B (2010): Retest reliability of reward-related BOLD signals. *Neuroimage* 50:1168–1176.
- Friston KJ (1999): Schizophrenia and the disconnection hypothesis. *Acta Psychiatr Scand Suppl* 395:68–79.
- Friston KJ, Frith CD (1995): Schizophrenia: A disconnection syndrome? *Clin Neurosci* 3:89–97.
- Friston K, Holmes A, Poline J, Grasby P, Williams S, Frackowiak R, Turner R (1995): Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
- Friston K, Harrison L, Penny W (2003): Dynamic causal modeling. *Neuroimage* 19:1273–1302.
- Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2007): Variational free energy and the Laplace approximation. *Neuroimage* 34:220–234.
- Frässle S, Paulus FM, Krach S, Schweinberger SR, Stephan KE, Jansen A (2015a): Mechanisms of hemispheric lateralization: Asymmetric interhemispheric recruitment in the face perception network. *Neuroimage* 124:977–988.
- Frässle S, Stephan KE, Friston KJ, Steup M, Krach S, Paulus FM, Jansen A (2015b): Test-retest reliability of dynamic causal modeling for fMRI. *Neuroimage* 117:56–66.
- Gorgolewski KJ, Storkey AJ, Bastin ME, Whittle I, Pernet C (2013): Single subject fMRI test-retest reliability metrics and confounding factors. *Neuroimage* 69:231–243.
- Grefkes C, Fink G (2012): Disruption of motor network connectivity post-stroke and its noninvasive neuromodulation. *Curr Opin Neurol* 25:670–675.
- Grefkes C, Eickhoff S, Nowak D, Dafotakis M, Fink G (2008): Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. *Neuroimage* 41:1382–1394.
- Grefkes C, Nowak D, Wang L, Dafotakis M, Eickhoff S, Fink G (2010): Modulating cortical connectivity in stroke patients by rTMS assessed with fMRI and dynamic causal modeling. *Neuroimage* 50:233–242.
- Grill-Spector K, Knouf N, Kanwisher N (2004): The fusiform face area subserves face perception, not generic within-category identification. *Nat Neurosci* 7:555–562.
- Grèzes J, Wicker B, Berthoz S, de Gelder B (2009): A failure to grasp the affective meaning of actions in autism spectrum disorder subjects. *Neuropsychologia* 47:1816–1825.
- Haxby J, Hoffman E, Gobbini M (2000): The distributed human neural system for face perception. *Trends Cogn Sci* 4:223–233.
- Huys Q, Moutoussis M, Williams J (2011): Are computational models of any use to psychiatry? *Neural Netw* 24:544–551.
- Ishai A (2008): Let's face it: It's a cortical network. *Neuroimage* 40:415–419.
- Jansen A, Menke R, Sommer J, Förster AF, Bruchmann S, Hempleman J, Weber B, Knecht S (2006): The assessment of hemispheric lateralization in functional MRI—Robustness and reproducibility. *Neuroimage* 33:204–217.
- Kanwisher N, McDermott J, Chun M (1997): The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17:4302–4311.
- Kass R, Steffey D (1989): Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Stat Assoc* 84:717–726.
- Kellermann T, Regenbogen C, De Vos M, Mößnang C, Finkelmeyer A, Habel U (2012): Effective connectivity of the human cerebellum during visual attention. *J Neurosci* 32:11453–11460.
- Kleinhans N, Richards T, Sterling L, Stegbauer K, Mahurin R, Johnson L, Greenon J, Dawson G, Aylward E (2008): Abnormal functional connectivity in autism spectrum disorders during face processing. *Brain* 131:1000–1012.
- Lee L, Friston K, Horwitz B (2006): Large-scale neural models and dynamic causal modelling. *Neuroimage* 30:1243–1254.
- 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.
- Madigan D, Raftery A (1994): Model selection and accounting for model uncertainty in graphical models using Occam's window. *JASA* 89:1535–1546.
- Maia T, Frank M (2011): From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci* 14:154–162.
- Mandeville J, Marota J, Ayata C, Zaharchuk G, Moskowitz M, Rosen B, Weisskoff R (1999): Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J Cereb Blood Flow Metab* 19:679–689.
- Mohr B, Landgrebe A, Schweinberger SR (2002): Interhemispheric cooperation for familiar but not unfamiliar face processing. *Neuropsychologia* 40:1841–1848.
- Monk CS, Weng SJ, Wiggins JL, Kurapati N, Louro HM, Carrasco M, Maslowsky J, Risi S, Lord C (2010): Neural circuitry of emotional face processing in autism spectrum disorders. *J Psychiatry Neurosci* 35:105–114.
- Montague P, Dolan R, Friston K, Dayan P (2012): Computational psychiatry. *Trends Cogn Sci* 16:72–80.
- Oldfield RC (1971): The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 9:97–113.
- Park H, Kim J, Lee S, Seok J, Chun J, Kim D, Lee J (2008): Corpus callosal connection mapping using cortical gray matter parcellation and DT-MRI. *Hum Brain Mapp* 29:503–516.
- Penny W, Roberts S (1999): Bayesian neural networks for classification: How useful is the evidence framework? *Neural Netw* 12:877–892.
- Penny WD, Stephan KE, Mechelli A, Friston KJ (2004): Modelling functional integration: A comparison of structural equation and dynamic causal models. *Neuroimage* 23:S264–S274.
- Penny W, Stephan K, Daunizeau J, Rosa M, Friston K, Schofield T, Leff A (2010): Comparing families of dynamic causal models. *PLoS Comput Biol* 6:e1000709.
- Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes AB, Sauer C, Tost H, Esslinger C, Colman P, Wilson F, Kirsch P, Meyer-Lindenberg A (2012): Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage* 60:1746–1758.

- Puce A, Allison T, Asgari M, Gore J, McCarthy G (1996): Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *J Neurosci* 16:5205–5215.
- Radulescu E, Minati L, Ganeshan B, Harrison NA, Gray MA, Beacher FD, Chatwin C, Young RC, Critchley HD (2013): Abnormalities in fronto-striatal connectivity within language networks relate to differences in grey-matter heterogeneity in Asperger syndrome. *Neuroimage Clin* 2:716–726.
- Reyt S, Picq C, Sinniger V, Clarençon D, Bonaz B, David O (2010): Dynamic Causal Modelling and physiological confounds: A functional MRI study of vagus nerve stimulation. *Neuroimage* 52:1456–1464.
- Roiser JP, Wigton R, Kilner JM, Mendez MA, Hon N, Friston KJ, Joyce EM (2013): Dysconnectivity in the frontoparietal attention network in schizophrenia. *Front Psychiatry* 4:176.
- Rowe J, Hughes L, Barker R, Owen A (2010): Dynamic causal modelling of effective connectivity from fMRI: Are results reproducible and sensitive to Parkinson's disease and its treatment? *Neuroimage* 52:1015–1026.
- Schlösser RG, Wagner G, Koch K, Dahnke R, Reichenbach JR, Sauer H (2008): Fronto-cingulate effective connectivity in major depression: A study with fMRI and dynamic causal modeling. *Neuroimage* 43:645–655.
- Schultz RT (2005): Developmental deficits in social perception in autism: The role of the amygdala and fusiform face area. *Int J Dev Neurosci* 23:125–141.
- Schuyler B, Ollinger J, Oakes T, Johnstone T, Davidson R (2010): Dynamic causal modeling applied to fMRI data shows high reliability. *Neuroimage* 49:603–611.
- Schweinberger SR, Baird LM, Blümner M, Kaufmann JM, Mohr B (2003): Interhemispheric cooperation for face recognition but not for affective facial expressions. *Neuropsychologia* 41:407–414.
- Seghier ML, Zeidman P, Neufeld NH, Leff AP, Price CJ (2010): Identifying abnormal connectivity in patients using dynamic causal modeling of fMRI responses. *Front Syst Neurosci* 4:142.
- Shrout P, Fleiss J (1979): Intraclass correlations—Uses in assessing rater reliability. *Psychol Bull* 86:420–428.
- Siman-Tov T, Mendelsohn A, Schonberg T, Avidan G, Podlipsky I, Pessoa L, Gadoth N, Ungerleider L, Hendler T (2007): Bihemispheric leftward bias in a Visuospatial attention-related network. *J Neurosci* 27:11271–11278.
- Specht K, Willmes K, Shah NJ, Jäncke L (2003): Assessment of reliability in functional imaging studies. *J Magn Reson Imaging* 17:463–471.
- Stephan KE (2004): On the role of general system theory for functional neuroimaging. *J Anat* 205:443–470.
- Stephan K, Mathys C (2014): Computational approaches to psychiatry. *Curr Opin Neurobiol* 25:85–92.
- Stephan K, Baldeweg T, Friston K (2006): Synaptic plasticity and dysconnection in schizophrenia. *Biol Psychiatry* 59:929–939.
- Stephan K, Marshall J, Penny W, Friston K, Fink G (2007): Interhemispheric integration of visual processing during task-driven lateralization. *J Neurosci* 27:3512–3522.
- Stephan K, Kasper L, Harrison L, Daunizeau J, den Ouden H, Breakspear M, Friston K (2008): Nonlinear dynamic causal models for fMRI. *Neuroimage* 42:649–662.
- Stephan K, Penny W, Daunizeau J, Moran R, Friston K (2009): Bayesian model selection for group studies. *Neuroimage* 46:1004–1017.
- Van Essen DC, Newsome WT, Bixby JL (1982): The pattern of interhemispheric connections and its relationship to extrastriate visual areas in the macaque monkey. *J Neurosci* 2:265–283.
- Vissers ME, Cohen MX, Geurts HM (2012): Brain connectivity and high functioning autism: A promising path of research that needs refined models, methodological convergence, and stronger behavioral links. *Neurosci Biobehav Rev* 36:604–625.
- Weigelt S, Koldewyn K, Kanwisher N (2012): Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neurosci Biobehav Rev* 36:1060–1084.
- Wilke M, Lidzba K (2007): LI-tool: A new toolbox to assess lateralization in functional MR-data. *J Neurosci Methods* 163:128–136.
- Woolrich MW, Stephan KE (2013): Biophysical network models and the human connectome. *Neuroimage* 80:330–338.
- Worsley K, Friston K (1995): Analysis of fMRI time-series revisited again. *Neuroimage* 2:173–181.
- Zeki S (1970): Interhemispheric connections of prestriate cortex in monkey. *Brain Res* 19:63.