# Interrater and Intermethod Reliability of Default Mode Network Selection

**Alexandre R. Franco,[1,2] Aaron Pritchard,[3] Vince D. Calhoun,[1,2] and Andrew R. Mayer[1,4]***

[1]*The Mind Research Network, Albuquerque, New Mexico*
[2]*Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, New Mexico*
[3]*Department of Student Research, University of New Mexico School of Medicine, Albuquerque, New Mexico*
[4]*Department of Neurology, University of New Mexico School of Medicine, Albuquerque, New Mexico*

**Abstract:** There has been a growing interest in the neuroimaging community regarding resting state data (i.e., passive mental activity) and the subsequent activation of the so-called default mode network (DMN). Although this network was originally characterized by a pattern of deactivation during active cognitive states, more recent applications of data-driven techniques such as independent component analysis (ICA) have permitted the analysis of brain activation during extended periods of truly passive mental activity. However, ICA requires the resultant components to be evaluated for "goodness of fit" via either human raters or more automated techniques. To our knowledge, an investigation on the reliability of either technique in determining the component that best corresponds to default-mode activity has not been performed. Moreover, it is not clear how automated techniques, which are necessarily dependent upon a template mask, are affected by the structures used to compose the mask. The current study investigated both interrater (human-human) reliability and intermethod (human-machine) reliability for determining DMN activation in 42 healthy controls. Results indicated that near perfect interrater reliability was achieved, whereas intermethod reliability was only within the moderate range. The latter was significantly improved via a weighted combination of the anterior and posterior cingulate nodes of the DMN. Implications for fully automating the component selection process are discussed. *Hum Brain Mapp 30:2293–2303, 2009.* 

**Key words:** fMRI; default mode network; independent component analysis; reliability

## INTRODUCTION

With the advent of advanced neuroimaging techniques such as functional magnetic resonance imaging (FMRI) and the arrival of increasingly sophisticated methods for interpreting these data, the study of the human brain at rest has flourished in recent years [Beckmann et al., 2005; Binder et al., 1999; Biswal et al., 1995; Buckner et al., 2008; Damoiseaux et al., 2006; Fox and Raichle, 2007; Gilbert et al., 2007; Greicius et al., 2003; Gusnard and Raichle, 2001; Jafri et al., 2007; Mason et al., 2007; Morcom and Fletcher, 2007; Raichle and Snyder, 2007; Raichle et al.,

Contract grant sponsor: The Mind Research Network—Mental Illness and Neuroscience Discovery; Contract grant numbers: DOE-DE-FG02-99ER62764, NIBIB R01 EB 000840.

*Correspondence to: Andrew Mayer, Ph.D., The Mind Research Network, Pete and Nancy Domenici Hall, 1101 Yale Blvd. NE, Albuquerque, NM 87131, USA. E-mail: amayer@mrn.org

Received for publication 14 April 2008; Revised 15 August 2008; Accepted 26 August 2008

DOI: 10.1002/hbm.20668

Published online 10 February 2009 in Wiley InterScience (www.interscience.wiley.com).

2001; Van de Ven et al., 2004]. In particular, there has been increasing interest in the concept of a default mode network (DMN), which exhibits coherent fluctuations during passive mental activity [Buckner et al., 2008], becomes deactivated during more demanding cognitive tasks [Binder et al., 1999; Gusnard and Raichle, 2001; Shulman et al., 1997], and which may serve as an intrinsic baseline state [Raichle and Gusnard, 2005]. Although the existence of the DMN has been investigated using multiple imaging modalities including FMRI, PET, and EEG [Greicius et al., 2003; Laufs et al., 2006; Raichle et al., 2001], controversy remains regarding the significance and applicability of the DMN model from both conceptual and quantitative perspectives [Gilbert et al., 2007; Mason et al., 2007; Morcom and Fletcher, 2007]. Specifically, several important questions remain regarding the consistency and reliability of the DMN over time, the analytic techniques used to characterize the brain at rest, and how reliable human and automated methods are in selecting DMN activity [Morcom and Fletcher, 2007]. The current investigation was conducted to rigorously address these important methodological concerns and to develop standardized methods that can be used to identify the DMN during the collection of resting state FMRI data.

Originally, the DMN was defined in terms of regions that were shown to be deactivated (hereafter referred to as task-induced deactivations) during periods of demanding mental activity [Binder et al., 1999; McKiernan et al., 2003, 2006; Shulman et al., 1997]. However, the magnitude of task-induced deactivation has been linked to the cognitive demands of tasks [Esposito et al., 2006; McKiernan et al., 2003], suggesting that task-induced deactivations may not be an ideal methodology for probing true passive mental activity. The identification of DMN has more recently been facilitated by the application of independent component analysis (ICA) to more extended periods (e.g., minutes rather than seconds) of resting state FMRI data, which permits the blind-source separation of overlapping signals into individual spatial or temporal components [Beckmann and Smith, 2004; Calhoun and Adali, 2006; Calhoun et al., 2007; McKeown et al., 1998, 2003; Van de Ven et al., 2004]. The data-driven nature of ICA permits the identification of networks of activity that occur when the brain is truly at rest for prolonged periods of time rather than examining the brain as it "deactivates" from a task. Voxels and regions that exhibit similar fluctuations in signal are subsequently grouped into different components, which are thought to represent different networks of correlated activity [Damoiseaux et al., 2006; De Luca et al., 2006]. However, a potential weakness of this data-driven approach is that the resulting components have no specific order and, therefore, must be evaluated and "selected" based on some preestablished criteria.

To date, the majority of researchers have relied on human observers to pick the component that best represents the DMN in ICA studies of the resting state [Beckmann et al., 2005; Damoiseaux et al., 2006; De Luca et al., 2006]. Although this system is practical, unresolved questions remain regarding interrater reliability both within studies and across different institutional sites. More importantly, manual selection may be subject to human error and is also a time consuming process in which all components must be visualized and then carefully analyzed prior to selecting the representative DMN component. An alternative approach is to spatially correlate the individual [Calhoun et al., 2007; Van de Ven et al., 2004] components with an ideal DMN template and then automatically select the component with the highest correlation [Calhoun et al., 2007; Greicius and Menon, 2004; Greicius et al., 2004; Van de Ven et al., 2004]. The template may be determined from the data collected during the study from a healthy group [Greicius and Menon, 2004; Greicius et al., 2004;) or based on the selection of regions from a stereotaxic atlas [Calhoun et al., 2007; Van de Ven et al., 2004]. Of these two potential methods, a template based on a stereotaxic atlas is preferable to increase reliability across different institutional sites.

However, determining the regions that constitute the DMN is a prerequisite for developing a spatial template (i.e., mask), which can subsequently be used to automate and standardize the component selection process across different institutions [Damoiseaux et al., 2006; Fox et al., 2005; Fransson, 2005; Greicius et al., 2003; Mazoyer et al., 2001; Raichle et al., 2001; Shulman et al., 1997]. Areas (see Table I) that have been consistently reported to constitute the DMN include the posterior cingulate (BA 23 and 31), posterior parietal (BA 7, 39, and 40), dorsolateral and superior frontal (BA 8, 9, and 10), and anterior cingulate (BA 11 and 32) cortex [Buckner et al., 2008]. Of these regions, the cingulate cortices may be particularly important for default-mode activity. Specifically, the posterior cingulate gyrus has been associated with high resting state metabolism in the DMN [Gusnard and Raichle, 2001] and previous studies have typically used seeds from the cingulate cortex to define the remainder of the DMN [Greicius et al., 2003]. Other researchers have reported DMN activity in the inferior temporal lobes (BA 19 and 37) [Damoiseaux et al., 2006; Fox et al., 2005; Mazoyer et al., 2001; Raichle et al., 2001; Shulman et al., 1997] although a recent review [Buckner et al., 2008] suggests these temporal areas may be less robust than the regions reported above. Therefore, some inconsistencies still remain in determining the neuronal regions that constitute the DMN.

To our knowledge, the reliability of automated versus human selection methods has not been investigated, nor have the instances in which the two methods of DMN component selection may disagree. To address these limitations in the current DMN literature, three aims related to the DMN selection process in the framework of a data-driven analysis are outlined. Two trained raters were asked to independently identify the DMN in 42 subjects following ICA analysis and rate their confidence in the selected component. The individual raters were then compared against each other and also against an automated component selection routine based on an atlas-derived template [Calhoun et al., 2007; Van de Ven et al., 2004].

**TABLE I. Commonly reported regions of DMN**

| Regions identified | BA | Publication (first author) |
|---|---|---|
| Anterior cingulate cortex | 11/32 | Shulman et al., 1997; Raichle et al., 2001; Mazoyer et al., 2001; Greicius et al., 2003; Fransson, 2005; Damoiseaux et al., 2006 |
| Dorsolateral and superior frontal gyrus | 8/9/10 | Shulman et al., 1997; Raichle et al., 2001; Mazoyer et al., 2001; Greicius et al., 2003; Fransson, 2005; Fox et al., 2005 |
| Inferior frontal cortex | 47 | Shulman et al., 1997; Mazoyer et al., 2001; Fransson, 2005 |
| Posterior cingulate cortex | 23/31 | Shulman et al., 1997; Raichle et al., 2001; Mazoyer et al., 2001; Greicius et al., 2003; Fransson, 2005; Fox et al., 2005; Damoiseaux et al., 2006 |
| Posterior parietal lobule | 7/39/40 | Shulman et al., 1997; Raichle et al., 2001; Mazoyer et al., 2001; Greicius et al., 2003; Fransson, 2005; Fox et al., 2005; Damoiseaux et al., 2006 |
| Inferior temporal gyrus | 19/37 | Shulman et al., 1997; Raichle et al., 2001; Mazoyer et al., 2001; Fox et al., 2005; Damoiseaux et al., 2006 |
| Parahippocampal gyrus | 30/36 | Greicius et al., 2003; Fransson, 2005; Fox et al., 2005; Damoiseaux et al., 2006 |

This method was designed to satisfy the first aim of the current study, which was to investigate both the reliability between two trained individuals (i.e., interrater reliability) and the reliability between human and more automated selection techniques (i.e., intermethod reliability). Our second aim was to examine how the length of data acquisition affected the quality of the DMN selection across both human and automated techniques. We predicted that both the human confidence ratings and magnitude of the correlation coefficient would improve as a function of data collection time. Our final aim was to investigate how altering the neuronal nodes that compromised the DMN mask would affect the subsequent reliability between automated and human methods. We predicted that a standard mask, comprised of the most commonly reported nodes of the DMN, would outperform both a sparser (anterior and posterior cingulate cortex only) and more complete (standard plus inferior temporal cortex) mask in terms of reliability and spatial correlation magnitude.

## METHODS

### Subjects

Data from 42 (15 females, mean age = 31.3 ± 9.6 years) subjects were examined in the current study. Subject data were pooled across two different experiments that measured different aspects of selective attention; however, the extended rest task was identical in both experiments and collected in separate runs (see Task section). Potential subjects with a history of neurological disease, major psychiatric disturbance, substance abuse, or psychoactive prescriptive medications were excluded from the study. Written informed consent was obtained from all participants prior to data collection, according to institutional guidelines at the University of New Mexico.

### Task

All data were collected on a Siemens Sonata 1.5 Tesla scanner. Subjects rested supine in the scanner with their head secured by chin and forehead straps, with additional foam padding to limit head motion within the head coil. Presentation software (Neurobehavioral Systems) was used for stimulus presentation and synchronization of stimulus events with the MRI scanner. Visual stimuli were rear-projected using a Sharp XG-C50X LCD projector on an opaque white Plexiglas projection screen.

The resting state data in the current study were pooled from two separate experiments. In both studies, alternating runs of complex attentional tasks (two runs for Study 1 and three for Study 2) and resting state data (one run) were collected. During both complex attention tasks, subjects ignored information from one sensory modality while performing a task based on information presented in an opposite sensory modality. In Study 1 ($N$ = 18), subjects rhythmically tapped in time to either a visual (reversing checkerboard) or auditory (pure tone) metronome while ignoring a simultaneously occurring crossmodal distractor. In Study 2 ($N$ = 24), subjects were asked to perform a numeric Stroop task in which they identified a target number presented in the visual or auditory modality while ignoring a crossmodal distractor. The results from these separate attention tasks will be presented in other publications.

The resting state task was identical in both experiments. Specifically, subjects were asked to relax and passively stare at a fixation cross (visual angle = 1.54°) for 3 min with their eyes open. Although research suggests that the resulting patterns of brain activity are similar for paradigms in which subjects keep their eyes opened or closed [Fransson, 2005; Raichle et al., 2001], subjects were requested to keep their eyes open to minimize the likelihood that they would fall asleep and to decrease the electrophysiological spectrum changes associated with sleep [Laufs et al., 2006]. The sequence of complex attentional and rest runs were repeated three times, resulting in three resting runs collected for a total of 9 min.

### MR Imaging

At the beginning of the scanning session, high resolution T1 [TE (echo time) = 4.76 ms, TR (repetition time) = 12 ms,

20° flip angle, number of excitations (NEX) = 1, slice thickness = 1.5 mm, FOV (field of view) = 256 mm, resolution = 256 × 256] anatomic images were collected. For each rest imaging series, 90 echoplanar images were collected using a single-shot, gradient-echo echoplanar pulse sequence [TR = 2000 ms; TE = 36 ms; flip angle = 90°; FOV = 256 mm; matrix size = 64 × 64]. The first two images of each run were eliminated to account for T1 equilibrium effects, leaving a total of 88 images for the final analyses. Twenty-eight contiguous sagittal 5-mm (Study 1) or 5.5-mm (Study 2) thick slices were selected to provide whole-brain coverage (voxel size: $4 \times 4 \times 5\ mm^3$ or $4 \times 4 \times 5.5\ mm^3$).

## Data Analyses

Functional images were generated and processed using a mixture of freeware and commercial packages including the Analysis of Functional NeuroImages (AFNI) [Cox, 1996], GIFT [Calhoun, 2004], MATLAB (Mathworks, Sherborn, MA) and FSL [Smith et al., 2004] packages. Time series images were first spatially registered (to the third image from the first resting run) in both two- and three-dimensional space to minimize effects of head motion, temporally interpolated to correct for slice-time acquisition differences, de-spiked, linearly detrended and spatially blurred using a 10-mm Gaussian full-width half-maximum filter.

The GIFT software package was then used to calculate the individual components on a subject-by-subject basis. Minimum description length (MDL) was used to establish the number of components necessary to be generated [Calhoun et al., 2001; Rissanen, 1983]. The ideal number of components ranged from 7 to 20 across the 42 subjects; therefore, 20 components were generated for all subjects in order to maintain consistency[1]. Components were calculated for each subject for 3 (corresponding to the first run), 6 (corresponding to the first two runs), and 9 (corresponding to all three runs) minutes of data collection using the Infomax algorithm [Bell and Sejnowski, 1995]. For the first 3 min of data collection, single-subject single-run ICA was performed. To calculate the components for 6 and 9 min, a group ICA [Calhoun et al., 2001] was implemented, as a simple concatenation of the time-courses of the individual runs cannot be performed since no baseline of resting state exists. The resulting components from individual and multirun ICAs were then converted to a 1 $mm^3$ standard stereotaxic coordinate space [Talairach and Tournoux, 1988].

## Rater Classification

The two human raters (A.F. and A.P.) initially trained on a subset of 10 randomly selected, de-identified datasets. The raters independently selected a single component from the 20 choices that best resembled the previously identified DMN, compared and then discussed their results. This training process was repeated twice.

The raters then independently selected a single component (from the 20 choices) that best resembled the previously identified DMN for each of the three time intervals for all of the subjects. Raters also assigned a confidence/quality rating to the component that was selected as most representative of DMN. A Likert scale of 1–5 was utilized with 1 indicating that the rater was "highly confident" that their selection reflected the DMN and 5 indicating that the rater was "not confident" (1 = Highly confident; 2 = Confident; 3 = Uncertain; 4 = Questionable; 5 = Not Confident). In addition, the raters also identified a second component that exhibited similar characteristics to the DMN, but was not as strong as their first choice. The raters selected the DMN components based on their spatial pattern of activation and based on the associated component time sequence, as the hemodynamic response of the DMN is defined by a slow fluctuating pattern, typically oscillating below 0.1 Hz [Biswal et al., 1995; Cordes et al., 2001; Fransson, 2005]. Interrater reliability for the first choice component was assessed using Cohen's Kappa Coefficient [Cohen, 1960]. A bivariate correlational analysis was also performed to assess agreement between the individual rater's confidence levels.

Following the independent rater assessments, a single DMN component was selected. If both raters agreed on a single DMN component for each subject/interval, no additional steps were taken. For the subjects/intervals in which there was disagreement, the two raters co-jointly examined all components and subsequently selected a consensus DMN component. Therefore, for each subject at each time interval, a single consensus component was identified that most closely resembled the DMN network. This component was then compared with the results of the spatial correlation analyses in the intermethod reliability analyses.

## Spatial Correlation

An automated method of DMN component selection was then performed based on the magnitude of the spatial correlation of each component with a series of DMN template masks, which were constructed using the Wake Forest University Pick atlas [Lancaster et al., 2000; Maldjian

[1]A separate analysis was conducted to determine if using the number of components specified by MDL on a subject-by-subject basis would significantly alter current results. Specifically, the MDL for each subject was first calculated for the 9-min dataset. The subsequent number of output components from the ICA was then restricted based on the MDL specifications (hereafter referred to as MDL analysis). Resulting components were then correlated with the standard mask to determine the component with the highest spatial correlation and compared with similar values obtained when components were fixed ($N = 20$) across all subjects (hereafter referred to as the fixed analysis). Results from a $t$-test indicated that there was no statistical difference ($t_{41} = -1.30$; $P = 0.20$) in the magnitude of the $z$-transformed spatial correlation coefficient between the MDL and fixed analyses.

et al., 2003]. Given the variability in findings across previous DMN studies, three separate DMN template masks were created to determine how the magnitude of the spatial correlation and intermethod reliability varied as a function of the mask.

### *Standard mask*

This mask was created by selecting the anatomical regions that have been most commonly reported to comprise the DMN network in previously published peer-reviewed research. The labels from the Wake Forest Atlas that constituted this mask included posterior cingulate (BAs 23/31), posterior parietal lobes (BAs 7/39/40), superior frontal gyrus (BAs 8/9/10), and anterior cingulate cortex (BAs 11/32) (see Table I; see Fig. 1).

### *Standard plus mask*

This mask included the inferior temporal gyrus (BAs 19/37) in addition to the labels from the standard mask.

### *Cingulate only mask*

Finally, a cingulate only mask was constructed by selecting only the anterior (BAs 11/32) and posterior (BAs 23/31) cingulate from the atlas.

The three resulting masks were then blurred using a 10-mm Gaussian full-width half-maximum filter to match the input data. A bivariate correlation analysis was then performed between each of the masks and the 20 components from the ICA to assess spatial correspondence. The component that resulted in the maximal correlation with each of the masks was then auto-selected to represent the DMN network. Resulting r scores were then transformed to a Fischer's $z$ to be used in all subsequent analyses.

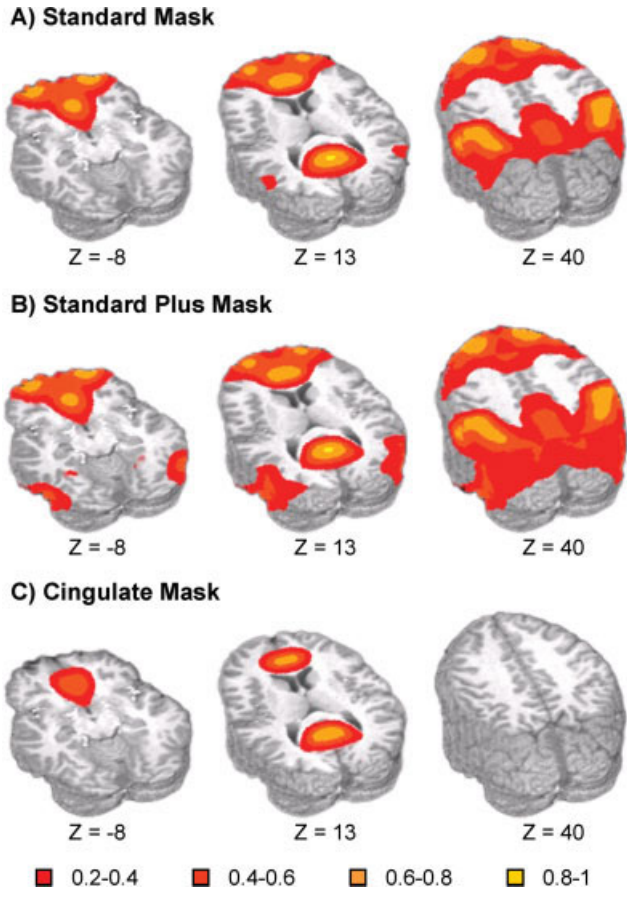


**Figure 1.**
Cartoon depiction of the (**A**) standard mask, (**B**) standard plus mask, and (**C**) cingulate only mask that were used in the principal analyses to determine magnitude of spatial correlation. All regions were selected according to the Wake Forest Pick Atlas as described in the methods. The colors of the masks correspond to the weights assigned to each voxel following the spatial blur, and therefore range between 0 and 1. Axial slice locations (Z) are presented according to the origin in Talairach space.

## RESULTS

The first set of analyses assessed whether there were any differences in the degree of mean rater confidence or spatial correlation with the standard mask across the two different attention studies and whether these variables significantly changed as a function of data acquisition time. Specifically, a 2 × 3 mixed ANOVA was first performed to determine if study (Study 1, Study 2) or run (Run 1, Run 2, Run 3) differentially affected the maximal spatial correlation obtained with the standard mask. Results indicated that the main effects of study and run, as well as the study × run interaction, were not significant ($P > 0.10$). Next, two 2 × 3 repeated-measures ANOVAs with study experiment (Study 1, Study 2) as a between-subjects factor and time of data collection (3, 6, and 9 min) as the within-subject factor were performed to investigate differences in the confidence rating and the maximal spatial correlation with the standard mask. Bivariate correlations indicated that there were significant correlations between the two raters at each epoch ($r$s ranging from 0.599 to 0.779; all $P$s $< 0.001$), suggesting that the mean rater confidence interval was an appropriate measure. For the analyses examining confidence rating (see Fig. 2 for selected DMN components corresponding to each confidence rating), only the main effect of time was significant ($F_{2,80} = 13.0$, $P < 0.001$). Follow-up, paired samples $t$-tests indicated that confidence ratings were better for both the 6- (mean = 2.69; $t_{41} = 3.4$, $P < 0.001$) and 9-min (mean = 2.61; $t_{41} = 7.1$, $P < 0.001$) scans compared with the 3 min (mean = 3.06) scan. There were no significant differences in confidence ratings between 6 and 9 min ($P > 0.10$). Results from the correlation analyses (see Fig. 3 for stratified presentation of DMN components according to spatial correlations) with the

standard mask indicated a main effect for time ($F_{2,80}$ = 18.1, $P < 0.001$), with a nonsignificant trend in the study by time interaction ($F_{2,80}$ = 3.1, $P$ = 0.06). Paired samples $t$-tests indicated that $z$-transformed correlation coefficients were greater for both the 6- (mean = 0.349; $t_{41}$ = −4.1, $P < 0.001$) and 9-min (mean = 0.355; $t_{41}$ = −5.6, $P < 0.001$) scan compared with the 3-min (mean = 0.315) scan, with no significant differences between 6 and 9 min scans ($P > .10$). Because effects associated with study were not significant in any of these analyses, this variable was excluded from further processing.

Kappa (κ)-coefficients were calculated to assess interrater reliability between the two human assessors for all 840 components for each of the three time intervals (see Table II). Results indicated that that there was almost perfect agreement [Cohen, 1960] for 3 (κ = 0.825; SE = 0.065), 6 and 9 min (both κ = 0.850; SE = 0.060 and SE = 0.062, respectively) of data collection between the two human raters. Three tests using the weighted least-squares (WLS) approach for comparing correlated κs [Barnhart and Williamson, 2002] were performed to assess whether the κ-coefficients were significantly different across the different acquisition lengths. However, there were no statistical differences ($P > 0.10$) between the different interrater κ-values as a function of time. For 100% of the cases in which interrater reliability was not achieved, the first-choice component selected by rater A was the second-choice component selected by rater B, or vice versa.

Similarly, κ-coefficients were calculated to assess the intermethod reliability between the automated component selection (spatial correlation) and the consensus component (see Table II). Results indicated that at 3 min of data collection there was only moderate agreement [Cohen, 1960] for the intermethod reliability with the standard mask (κ = 0.549; SE = 0.083), the standard plus mask (κ = 0.473; SE = 0.083), and the cingulate only mask (κ = 0.574; SE = 0.081). The κ-coefficients across the three masks were not statistically different by the WLS method ($P > 0.10$). At 6 min of data collection, moderate agreement was obtained for the standard mask (κ = 0.449; SE = 0.085) and the standard plus mask (κ = 0.499; SE = 0.084), whereas substantial agreement was achieved for the cingulate only mask (κ = 0.674; SE = 0.080). WLS testing indicated that the κ-coefficient for the cingulate only mask was significantly higher compared with the standard mask at 6 min of data collection ($P < 0.05$). Finally, moderate agreement was achieved for the standard (κ = 0.449; SE = 0.084) and cingulate only (κ = 0.574; SE = 0.071) masks at 9 min of data collection; however, only fair agreement [Cohen, 1960] was achieved for the standard plus masks

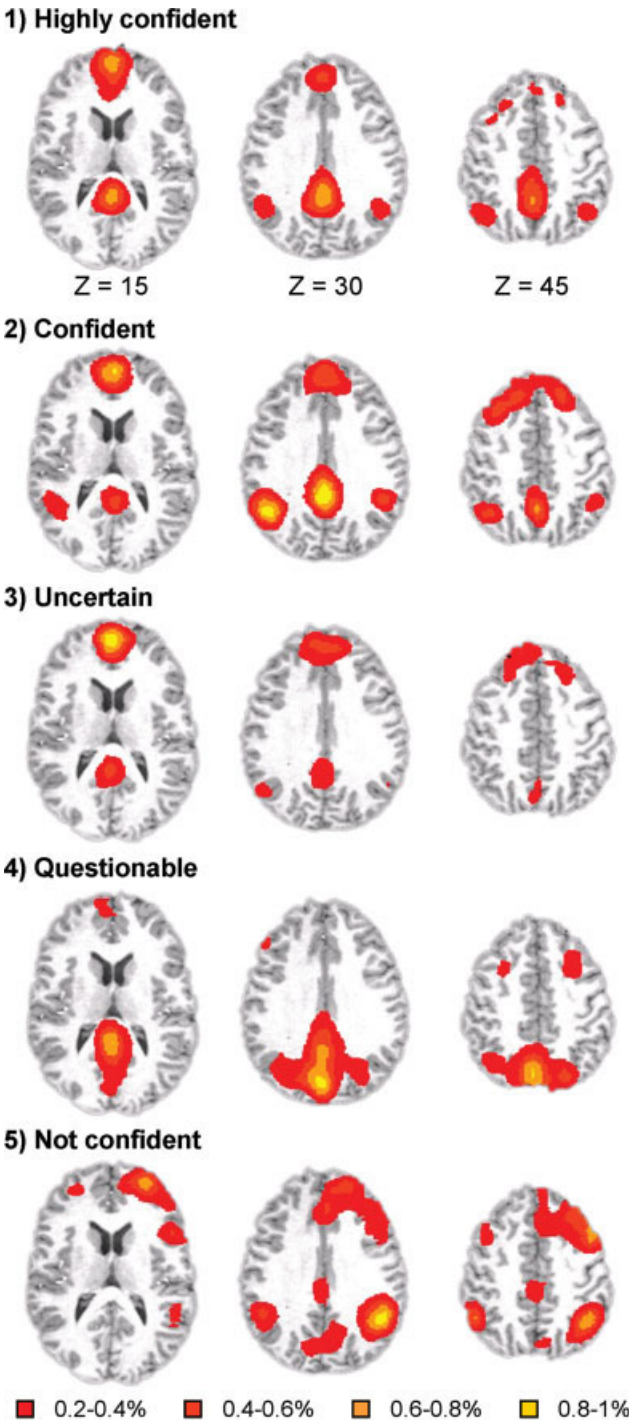


**Figure 2.**

Examples of five individual subject's consensus DMN component from the 9-min analyses. Components are displayed according to the confidence rating assigned by both human raters on a 5-point Likert scale (1 = Highly confident; 2 = Confident; 3 = Uncertain; 4 = Questionable; 5 = Not Confident). Following independent component analyses, the resulting component values are not on any scale. Therefore, for display purposes and consistency across subjects, these component maps were scaled and thresholded based on the maximum voxel-wise value of the component (see legend). Axial slice locations ($Z$) are presented according to the origin in Talairach space.

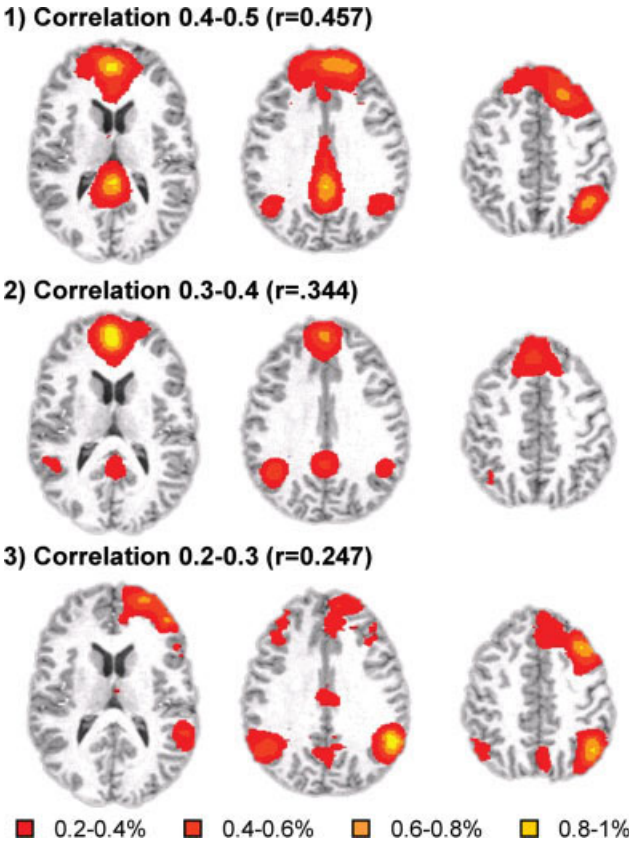


**Figure 3.**

Examples of three individual subject's DMN components that were selected (i.e., maximum correlation) during the automated procedure using the standard mask. Selected components were stratified into three categories according to spatial correlation ranges, and the exact value of the selected component is presented in parenthesis. For display purposes and consistency across subjects, resultant component maps were scaled and thresholded based on the maximum voxel-wise value of the component (see legend). Axial slice locations (Z) are identical to those presented in Figure 2.

for the same time period. There were no statistical differences ($P > 0.10$) in the magnitude of the κ-coefficients across the three different masks.

A total of nine WLS tests (three at each time period) were also performed to compare the κ-coefficients between the interrater and the intermethod reliability methods for selecting the DMN component. The interrater κ-coefficients were significantly ($P < 0.05$) higher compared with the intermethod reliability obtained using the standard and standard plus masks across the 3-, 6-, and 9-min time periods. The interrater κ-coefficients were also significantly higher ($P < 0.05$) than the intermethod reliability coefficients obtained using the cingulate only mask at 3 and 9 min. At 6 min of data collection, a nonsignificant trend ($P = 0.098$) was observed.

Next, 1 × 3 (standard, cingulate only, standard plus) repeated-measures ANOVAs were performed to compare the *z*-transformed correlation coefficients for the different masks for each of the three different periods of data collection. Results indicated a main effect of mask for 3 ($F_{2,82} = 6.4$, $P < 0.005$), 6 ($F_{2,82} = 12.5$, $P < 0.001$), and 9 ($F_{2,82} = 5.7$, $P < 0.005$) minutes of data collection. Follow-up *t*-tests indicated that the coefficients were greater for both the standard (mean = 0.316) compared with the standard plus (mean = 0.289) mask ($t_{41} = 10.7$, $P < 0.001$) and for cingulate only (mean = 0.327) compared with standard plus mask ($t_{41} = 3.0$, $P < 0.005$) at 3 min of data collection. There were no significant differences in the correlation coefficient between the standard and cingulate only mask ($P > 0.10$). For 6 min of data collection, the standard (mean = 0.349; $t_{41} = 9.6$, $P < 0.001$) and cingulate only (mean = 0.379; $t_{41} = 4.2$, $P < 0.001$) masks exhibited a higher coefficient than the standard plus mask (mean = 0.323). In addition, the spatial correlation was also significantly greater for the cingulate only compared with the standard mask ($t_{41} = -2.2$, $P < 0.05$). Finally, for 9 min of data collection, the standard (mean = 0.355; $t_{41} = 12.5$, $P < 0.001$) and cingulate only (mean = 0.369; $t_{41} = 2.8$, $P < 0.01$) masks exhibited higher coefficients than the standard plus mask (mean = 0.327) but did not statistically differ between each other ($P > 0.10$).

## Supplementary Analyses

Contrary to our predictions, current results indicated a slight improvement in performance (magnitude of spatial correlation and κ-coefficients) for the cingulate only mask compared with the standard mask. Therefore, differential

**TABLE II. κ Coefficients and agreement ratios for interrater and intermethod reliability**

| | 3 minutes | 6 minutes | 9 minutes | Average |
|---|---|---|---|---|
| Interrater | 0.825 k (AP) 35/42 Cases | 0.850 k (AP) 36/42 Cases | 0.850 k (AP) 36/42 Cases | 0.841 k (AP) |
| Intermethod (standard mask) | 0.549 k (M) 24/42 Cases | 0.449 k (M) 20/42 Cases | 0.449 k (M) 20/42 Cases | 0.482 k (M) |
| Intermethod (standard plus mark) | 0.473 k (M) 21/42 Cases | 0.499 k (M) 22/42 Cases | 0.398 k (F) 18/42 Cases | 0.467 k (M) |
| Intermethod (cingulate only mask) | 0.574 k (M) 25/42 Cases | 0.674 k (S) 29/42 Cases | 0.574 k (M) 25/42 Cases | 0.607 k (S) |
| Intermethod (standard weighted mask[a]) | 0.699 k (S) 30/42 Cases | 0.649 k (S) 28/42 Cases | 0.624 k (S) 27/42 Cases | 0.657 k (S) |

F, Fair agreement; M, moderate agreement; S, Substantial agreement; AP, Almost perfect agreement.
[a] The standard weighted mask was constructed to conduct supplementary analyses.

weighting parameters were assigned to different nodes of the standard mask to determine if the performance could be further improved in the automated DMN selection routine. Specifically, a weighting factor of 3 was assigned to the nodes corresponding to both the anterior and posterior cingulate gyrus, whereas the remaining nodes of the standard mask remained weighted at 1. This mask was then blurred with a 10-mm Gaussian kernel, and spatial correlations and intermethod reliability were recalculated. WLS tests indicated that the intermethod κ-coefficients were significantly higher ($P < 0.05$) for the standard weighted compared with the standard mask at 3 ($\kappa = 0.699$; SE = 0.076) and 6 ($\kappa = 0.649$; SE = 0.081) minutes of data collection, with a nonsignificant trend present for 9 ($\kappa = 0.624$; SE = 0.071) min of data collection ($P = 0.054$). When comparing interrater and intermethod reliability methods, the interrater reliability again outperformed the intermethod reliability with the standard weighted mask at both 6 and 9 min of data acquisition ($P < 0.05$), but no differences were noted at 3 min ($P > 0.10$).

A 2 × 3 repeated measures ANOVA [Mask (standard vs. standard weighted) × Time (3, 6, and 9 min)] indicated a significant effect of mask ($F_{1,41} = 34.2$, $P < 0.001$) and of time ($F_{1,41} = 37.2$, $P < 0.001$) for the magnitude of the *z*-transformed correlation coefficient. Follow-up *t*-tests indicated that the magnitude of the correlation was greater for the standard weighted (mean = 0.386) compared with standard (mean = 0.340) mask ($t_{41} = 5.84$, $P < 0.001$). The effects of time were similar to those reported for the standard mask and are not repeated here. In addition, a qualitative examination of the results from the standard mask analyses indicated that one of the three components with the highest spatial correlation corresponded to the consensus component selected by the human raters in 88.1% of the cases. For all of these components (126 components = 3 components per subject × 42 subjects), the majority of the power spectrum (above 50%) was concentrated below 0.1 Hz. In contrast, results from the standard weighted mask analyses indicated that one of the top three components always (100%) corresponded to the human selected consensus component.

Finally, the performance of the standard mask (derived from a stereotaxic atlas) was directly compared with the performance of a DMN template empirically derived from a subset of the subjects' data, as has been done in previous studies [Greicius and Menon, 2004; Greicius et al., 2004]. Specifically, the averaged DMN derived from Study 2 ($N = 24$) was used as a template mask for Study 1 ($N = 18$). Results indicated that the spatial correlation of the maximal component was significantly higher with the template mask (mean = 0.58987 ± 0.07729) compared with the stereotaxically created standard mask (mean = 0.32982 ± 0.04836; $t = 12.45$, $P < 0.001$). However, there was not any significant improvement in the intermethod reliability between the template mask (12 consistently identified components with empirically determined mask) and the standard mask (11 consistently identified components).

## DISCUSSION

The primary aim of the current experiment was to examine the reliability of two popular methods for detecting the component that was most representative of DMN activity during ICA analyses. Although several previous studies have utilized data-driven techniques to assess functional activity during the resting state, to date there has not been an assessment of how reliable DMN selection is using either manual selection [Beckmann et al., 2005; Damoiseaux et al., 2006; De Luca et al., 2006] or more automated methods [Calhoun et al., 2007; Greicius and Menon, 2004; Greicius et al., 2004; Van de Ven et al., 2004]. Results indicated that, although almost perfect agreement was reached between two highly trained manual raters (i.e., interrater reliability), agreement between automated routines (spatial correlation analyses) and manual raters (i.e., intermethod reliability) was only in the moderate range [Cohen, 1960]. To our knowledge, this is the first demonstration of the reliability of DMN activity by independent human raters during longer epochs of uninterrupted passive mental activity in a data-driven analysis technique. In the cases where the two human raters did not agree on their first choice component (~15% across all three intervals), there was 100% agreement among the top two choices. These results are consistent with previous findings of multiple spatially correlated networks during passive mental activity [Damoiseaux et al., 2006; De Luca et al., 2006] and suggest that more than one component may be associated with DMN activity.

Moreover, the quantity of time allocated to passive mental activity significantly impacted both indices of quality for DMN activity. Specifically, the magnitude of both the spatial correlation and rater confidence level significantly improved as a function of time from 3 min of data collection to 6 min of data collection. However, there were no substantial differences between the indices of quality at the 6- compared with 9-min intervals. Collectively, these results suggest that a qualitative and reliable assessment of passive mental activity can be obtained in approximately a 6-minute epoch of data collection. Determining the minimum amount of time necessary for the reliable assessment of DMN activity will be crucial for determining applicability of resting-state scans in clinical applications such as the diagnosis of Alzheimer's dementia [Greicius et al., 2004], where time is often a critical constraint in the collection of movement-free data.

Only moderate agreement [Cohen, 1960] was achieved between human and more automated routines for selecting the DMN based on a standard mask, with intermethod reliability occurring for only ~51% of the cases. Human raters were adopted as the ground truth in the current experiment as the majority of previous studies have relied on human intervention to select DMN activity during data-driven approaches [Beckmann et al., 2005; Damoiseaux et al., 2006; De Luca et al., 2006]. However, to date, a gold standard or computational formula for identifying

the DMN does not exist, which partially explains why DMN activity remains an active controversy in the field [Morcom and Fletcher, 2007]. Current findings suggest that methodological approaches for DMN selection are an important step in determining DMN activity. Moreover, current results also suggest that previous studies that relied on a spatial coefficient to select the DMN either by using a map derived from a stereotaxic atlas [Calhoun et al., 2007; Van de Ven et al., 2004] or a map empirically derived from an independent group of healthy controls [Greicius and Menon, 2004; Greicius et al., 2004] may not readily generalize to human selection methods. However, current results also indicated that the three most highly correlated components corresponded to the component selected by human raters in the majority of cases (88.1% with the standard mask and 100% with the standard weighted mask), suggesting that results from human and automated techniques may not be too discordant.

Current results also provide preliminary evidence that differentially weighting certain nodes of the DMN may increase intermethod reliability. Specifically, supplementary analyses indicated that a standard mask preferentially weighted toward the anterior and posterior cingulate gyrus achieved substantial intermethod agreement and was associated with higher magnitudes of the correlation coefficient compared with the standard mask. Previous researchers have utilized the anterior and posterior cingulate as seed voxels to detect the remainder of the DMN [Fox et al., 2005; Fransson, 2005; Greicius et al., 2003; Shulman et al., 1997] and resting-state metabolism has shown to be the highest in the posterior cingulate gyrus [Gusnard and Raichle, 2001]. Meta-analyses of task-induced deactivations during five separate tasks also identified the anterior cingulate gyrus as one of the only regions that demonstrated consistent deactivation [Wicker et al., 2003]. Finally, a connectivity analysis of the intrinsic correlations within the DMN demonstrated that the medial frontal and posterior cingulate regions may serve as a central hub for the DMN [Buckner et al., 2008]. These findings reinforce the idea that the anterior and posterior cingulate cortex may form the core of the DMN that is consistently activated in humans during undirected mental states [Buckner et al., 2008]. Moreover, our current finding of a lower spatial correlation with the standard plus mask is consistent with the suggestion that the medial temporal lobe system may form an interacting subsystem with the DMN [Buckner et al., 2008].

In spite of the lower intermethod κ-values obtained in the current study, the automatic or semiautomatic detection of the DMN remains an important goal for quantifying passive mental activity during rest for several reasons. Foremost, manual selection of the DMN requires extensive human resources. For example, in the current experiment each human rater was asked to visually evaluate 840 different components (42 subjects $\times$ 20 components), greatly increasing the likelihood of fatigue and error. Second is objectivity; an algorithm is typically more consistent and unbiased compared with human methods. This may be especially important for clinical applications in which the population of interest is too impaired to perform cognitive tasks, but where rater bias may play an important role in determining outcome measures (e.g., using DMN activity as diagnostic criteria for differentiating clinical from healthy populations). Indeed, DMN activity has already proved to be fertile ground for the advancement of the understanding of various areas of clinical research including schizophrenia [Calhoun et al., 2007; Garrity et al., 2007] and Alzheimer's disease [Greicius et al., 2004], suggesting neuroimaging studies of passive mental activity may eventually contribute to difficult differential diagnoses.

Current and previous findings suggest that a mixture of manual and automated techniques may provide the best approach for selecting the DMN component during data-driven analyses. For example, previous work [Damoiseaux et al., 2006] has demonstrated that the power spectrum of the DMN component may contain important features that could potentially be used to improve the selection process via machine learning algorithms. Specifically, the peak of the power spectrum of the DMN component usually occurs below 0.1 Hz [Biswal et al., 1995; Cordes et al., 2001; Fransson, 2005; Greicius et al., 2004]. In the current experiment, the majority (>50%) of the frequency distribution and the peak of the power spectrum were below 0.1 Hz for the top three correlated components. This temporal information could be then used to reduce the number of components that were entered into either an automated (e.g., spatial correlation) or manual selection process, thereby greatly reducing computational and human time.

Current results also demonstrated exact correspondence between the human consensus component and one of the three components with the highest correlation selected by the standard weighted mask. Therefore, when necessary to identify a single subject DMN, one could calculate the correlation of the components with a stereotaxically defined mask followed by human selection of one of the top three correlated components. This would drastically reduce the time necessary to manually select the DMN and greatly increase the objectivity of the manual selection process. In the current application, this approach would have reduced the number of components that required visual examination from 840 to 126, thus saving substantial human computational resources and reducing the likelihood of error.

Finally, a question remains regarding the use of a stereotaxic (generated through anatomical labels) or empirically (generated from DMN data) derived template as a spatial mask if semiautomatic or fully automatic techniques are to be adopted [Calhoun et al., 2007; Greicius and Menon, 2004; Greicius et al., 2004; Van de Ven et al., 2004]. In the current study, an empirically derived template mask significantly improved the spatial correlation coefficient compared with a stereotaxic mask but did not improve the intermethod reliability. This improvement in the correlation coefficient was expected given that the data were collected in a similar population under identical experimental

conditions. A more informative comparison would contrast the different masks' performance with data collected at different scanners at different institutions with different sample characteristics. However, current results do suggest that the observed magnitude of the spatial correlation is very dependent on the method used to generate the spatial template, which may have more importance for clinical studies. The benefit of a mask derived from a stereotaxic atlas is that it can be standardized and easily reproduced across different research institutions so that individual studies become less dependent on previous data, sample characteristics or manual selection of the DMN component.

In summary, the reliable identification of the DMN is a crucial first step for any study using a data-driven analysis technique to quantify brain activation during passive mental activity. Current results are the first to suggest that high levels of interrater reliability can be achieved during manual selection, whereas intermethod reliability (i.e., manual and automated selection routines) was only in the moderate range. Current results suggest that automated techniques may be useful as a first step for reducing the number of components followed by a more thorough evaluation by human raters. The use of a standardized template and methodology for automating DMN selection should reduce some of the methodological concerns regarding extended periods of passive mental activity while increasing the reproducibility of results across different sites. Future studies should determine whether weighting different nodes of the DMN or other data features can be utilized to increase the reliability of these automated techniques as was observed in the current study.

## ACKNOWLEDGMENTS

The authors thank Diana South for assistance with data collection, Charles Gasparovic, Ph.D. for help with technical support, and Edward J. Bedrick, Ph.D. for help with statistical calculations.

## REFERENCES

Barnhart HX, Williamson JM (2002): Weighted least-squares approach for comparing correlated kappa. Biometrics 58:1012–1019.

Beckmann CF, Smith SM (2004): Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans Med Imaging 23:137–152.

Beckmann CF, DeLuca M, Devlin JT, Smith SM (2005): Investigations into resting-state connectivity using independent component analysis. Philos Trans R Soc Lond B Biol Sci 360:1001–1013.

Bell AJ, Sejnowski TJ (1995): An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7:1129–1159.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Rao SM, Cox RW (1999): Conceptual processing during the conscious resting state. A functional MRI study. J Cogn Neurosci 11:80–95.

Biswal B, Yetkin FZ, Haughton VM, Hyde JS (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34:537–541.

Buckner RL, Andrews-Hanna JR, Schacter DL (2008): The brain's default network: Anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124:1–38.

Calhoun VD (2004): Group ICA of fMRI Toolbox (GIFT).

Calhoun VD, Adali T (2006): Unmixing fMRI with independent component analysis. IEEE Eng Med Biol Mag 25:79–90.

Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001): A method for making group inferences from functional MRI data using independent component analysis. Hum Brain Mapp 14:140–151.

Calhoun VD, Maciejewski PK, Pearlson GD, Kiehl KA (2007): Temporal lobe and "default" hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. Hum Brain Mapp (in press).

Cohen J (1960): A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–45.

Cordes D, Haughton VM, Arfanakis K, Carew JD, Turski PA, Moritz CH, Quigley MA, Meyerand ME (2001): Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. AJNR Am J Neuroradiol 22:1326–1333.

Cox RW (1996): AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173.

Damoiseaux JS, Rombouts SA, Barkhof F, Scheltens P, Stam CJ, Smith SM, Beckmann CF (2006): Consistent resting-state networks across healthy subjects. Proc Natl Acad Sci USA 103:13848–13853.

De Luca M, Beckmann CF, De SN, Matthews PM, Smith SM (2006): fMRI resting state networks define distinct modes of long-distance interactions in the human brain. Neuroimage 29:1359–1367.

Esposito F, Bertolino A, Scarabino T, Latorre V, Blasi G, Popolizio T, Tedeschi G, Cirillo S, Goebel R, Di Salle F (2006): Independent component model of the default-mode brain function: Assessing the impact of active thinking. Brain Res Bull 70:263–269.

Fox MD, Raichle ME (2007): Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8:700–711.

Fox MD, Snyder AZ, Vincent JL, Corbetta M, Van E, Raichle ME (2005): The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci USA 102:9673–9678.

Fransson P (2005): Spontaneous low-frequency BOLD signal fluctuations: An fMRI investigation of the resting-state default mode of brain function hypothesis. Hum Brain Mapp 26:15–29.

Garrity AG, Pearlson GD, McKiernan K, Lloyd D, Kiehl KA, Calhoun VD (2007): Aberrant "default mode" functional connectivity in schizophrenia. Am J Psychiatry 164:450–457.

Gilbert SJ, Dumontheil I, Simons JS, Frith CD, Burgess PW (2007): Comment on "Wandering minds: The default network and stimulus-independent thought". Science 317:43.

Greicius MD, Menon V (2004): Default-mode activity during a passive sensory task: Uncoupled from deactivation but impacting activation. J Cogn Neurosci 16:1484–1492.

Greicius MD, Krasnow B, Reiss AL, Menon V (2003): Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. Proc Natl Acad Sci USA 100:253–258.

Greicius MD, Srivastava G, Reiss AL, Menon V (2004): Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. Proc Natl Acad Sci USA 101:4637–4642.

Gusnard DA, Raichle ME (2001): Searching for a baseline: Functional imaging and the resting human brain. Nat Rev Neurosci 2:685–694.

Jafri M, Pearlson GD, Stevens M, Calhoun VD (2008): A method for functional network connectivity among spatially independent resting-state components in schizophrenia. Neuroimage 39:1666–1681.

Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT (2000): Automated Talairach atlas labels for functional brain mapping. Hum Brain Mapp 10:120–131.

Laufs H, Holt JL, Elfont R, Krams M, Paul JS, Krakow K, Kleinschmidt A (2006): Where the BOLD signal goes when alpha EEG leaves. Neuroimage 31:1408–1418.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003): An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, Macrae CN (2007): Wandering minds: The default network and stimulus-independent thought. Science 315:393–395.

Mazoyer B, Zago L, Mellet E, Bricogne S, Etard O, Houde O, Crivello F, Joliot M, Petit L, Tzourio-Mazoyer N (2001): Cortical networks for working memory and executive functions sustain the conscious resting state in man. Brain Res Bull 54:287–298.

McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, Sejnowski TJ (1998): Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp 6:160–188.

McKeown MJ, Hansen LK, Sejnowsk TJ (2003): Independent component analysis of functional MRI: What is signal and what is noise? Curr Opin Neurobiol 13:620–629.

McKiernan KA, Kaufman JN, Kucera-Thompson J, Binder JR (2003): A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. J Cogn Neurosci 15:394–408.

McKiernan KA, D'Angelo BR, Kaufman JN, Binder JR (2006): Interrupting the "stream of consciousness": An fMRI investigation. Neuroimage 29:1185–1191.

Morcom AM, Fletcher PC (2007): Does the brain have a baseline? Why we should be resisting a rest. Neuroimage 37:1073–1082.

Raichle ME, Gusnard DA (2005): Intrinsic brain activity sets the stage for expression of motivated behavior. J Comp Neurol 493:167–176.

Raichle ME, Snyder AZ (2007): A default mode of brain function: A brief history of an evolving idea. Neuroimage 37:1083–1090.

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL (2001): A default mode of brain function. Proc Natl Acad Sci USA 98:676–682.

Rissanen J (1983): A universal prior for integers and estimation by minimum description length. Ann Stat 11:416–431.

Shulman GL, Fiez JA, Corbetta M, Buckner RL, Miezin FM, Raichle ME, Petersen SE (1997): Common blood flow changes across visual tasks. II. Decreases in cerebral cortex. J Cogn Neurosci 9:648–663.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De LM, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De SN, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1):S208–S219.

Talairach J, Tournoux P (1988): Co-Planar Stereotaxic Atlas of the Human Brain. New York: George Thieme Verlag.

Van de Ven VG, Formisano E, Prvulovic D, Roeder CH, Linden DE (2004): Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Hum Brain Mapp 22:165–178.

Wicker B, Ruby P, Royet JP, Fonlupt P (2003): A relation between rest and the self in the brain? Brain Res Brain Res Rev 43:224–230.