NeuroImage 45 (2009) 758–768

Contents lists available at ScienceDirect

NeuroImage

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

Measuring fMRI reliability with the intra-class correlation coefficient

Alejandro Caceres [*], Deanna L. Hall, Fernando O. Zelaya, Steven C.R. Williams, Mitul A. Mehta

Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK

ARTICLE INFO

Article history:
Received 22 May 2008
Revised 10 November 2008
Accepted 17 December 2008
Available online 31 December 2008

Keywords:
fMRI
Reliability
Intra-class correlation coefficient
Working memory
Auditory cortex

ABSTRACT

The intra-class class correlation coefficient (ICC) is a prominent statistic to measure test–retest reliability of fMRI data. It can be used to address the question of whether regions of high group activation in a first scan session will show preserved subject differentiability in a second session. With this purpose, we present a method that extends voxel-wise ICC analysis. We show that voxels with high group activation have more probability of being reliable, if a subsequent session is performed, than typical voxels across the brain or across white matter. We also find that the existence of some voxels with high ICC but low group activation can be explained by stable signals across sessions that poorly fit the HRF model. At a region of interest level, we show that our voxel-wise ICC calculation is more robust than previous implementations under variations of smoothing and cluster size. The method also allows formal comparisons between the reliabilities of given brain regions; aimed at establishing which ROIs discriminate best between individuals. The method is applied to an auditory and a verbal working memory task. A reliability toolbox for SPM5 is provided at http://brainmap.co.uk.

Introduction

Test–retest studies are essential to determine the reliability of functional magnetic resonance imaging (fMRI). Together with studies of statistical power, they constitute the basis for the design of large longitudinal experiments. Previous test–retest studies have quantified fMRI reliability for a wide number of tasks, ranging from primary sensory (motor, visual and auditory) to cognitive and emotional paradigms (Liu et al., 2004; Yoo et al., 2005; Kong et al., 2007; Rombouts et al., 1997; Kiehl and Liddle, 2003; Manoach et al., 2001; Wei et al., 2004; Aron et al., 2006; Johnstone et al., 2005). Their results are varied as are the statistical methods used to report the analysis of repeated-measures. A prominent measure of reliability, amongst these, is the intra-class correlation coefficient (ICC), which informs on the ability of fMRI to assess differences in brain activity between subjects.

A number of specific methods to neuroimaging data have been developed to assess the stability of brain activation. An initial interest is to assess whether the volume of group activation in a first session is similar to that of a second session. Some studies (Yoo et al., 2005; Rombouts et al., 1997) focus on the extent of the activation, comparing the sessions by the amount of activated voxels in each occasion. The main weakness of this approach is that it strongly depends on the statistical threshold used to define activation. One can easily conceive a hypothetical situation where both group maps are identical except for an additive constant across the whole brain volume. In that situation the method may report low agreement, when there is in fact a consistent signal distribution. A further limitation is that, even in the case where the two activated volumes are the same, it does not inform whether each subject activated consistently within the group. The same group activation could be obtained by a fortuitous rearrangement of individual activations.

A better alternative is to determine the areas of group activation in the first session and then ask if, in the same region, the rank order of subject activations will be preserved in a subsequent session. Or equivalently, we can ask whether the level of group activation of the first session can predict the consistency in subject activations. These issues can be addressed with standard statistical analysis, the ICC being the most appropriate.

The most general approach would be, however, to assess the repeatability of observations by quantifying the error measurement, given by the within-subject variance (Zandbelt et al., 2008). Repeatable activations are those whose within-subject variances are smaller than an agreed limit (Bland and Altman, 1986). In fMRI there is not yet a predetermined standard for the acceptance of the error. And in consequence, reliability is more commonly assessed rather than repeatability.

Reliability is understood as a relative scaling of the measurement error. And although, it is usually interchanged with reproducibility, we reserve this term for the more fundamental case of experimental results being independent of the experimenter or the population sample.

Two different types of statistic can be regarded as the scaling of the measurement error. The first kind is the coefficient of variation (CV)

[*] Corresponding author. Institute of Psychiatry, PO 89, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. Fax: +44 203 228.
E-mail address: alejandro.caceres@iop.kcl.ac.uk (A. Caceres).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.12.035

where the error variance is scaled by the magnitude of activation. More precisely, the CV is the ratio between the standard deviation and the magnitude of signal change between two conditions. An example of its implementation to neuroimaging data is given by Tjandra et al. (2005), where they compare the CV for BOLD and MR perfusion imaging. The main limitation of the coefficient of variation, for the purpose of this work, is that it cannot be used to assess the relative error when the observation values are low or negative, even if the rank order of the subjects is preserved.

The second kind of scaling of the measurement error is the ICC (Shrout and Fleiss, 1979), defined as the ratio of the between-subject variance and the total variance. Given that the error variance is included in the total variance, in some cases, the ICC can be written in terms of the error variance divided by between-subject variance. The coefficient conveniently assesses either the absolute or consistent agreement of subject activations from session to session (McGraw and Wong, 1996). Intra-subject reliability ranges from zero (no reliability) to one (perfect reliability). As Bland and Altman (1996) explain, the coefficient can be understood as a measure of discrimination between subjects. In the context of neuroimaging data, it then allows the identification of single observations (e.g. voxel *t*-scores) to subjects and, therefore, the tracking of individuals across sessions.

A main feature of the ICC is that it is calculated from the variance structure of the data. Based on this characteristic, it has been used to show that the between-subject variance of BOLD activation is higher than the within-subject variance (Wei et al., 2004). A more recent study (Friedman et al., 2007) shows between-site reliability derived from a variance component analysis. Since the ICC depends exclusively on the variance, it can be computed for any level of activation. It can be shown (see Materials and methods section) that reliability brings additional and *complementary* information to group activation. In particular we can have situations in which voxels that fail a group *t*-test can present high reliability, meaning that their measurements are still consistent across sessions. For instance, non-linear responses that poorly fit the hemodynamic model may yet be consistent for each individual subject. A more fundamental question is to assess whether voxels of high group activation in the first session are likely to be reliable, or, if one-session group-activation is a predictor of intra-subject reliability.

There are three main possible ICC implementations on neuroimaging data, as reported in the literature. Typically, a summary statistic for each subject is obtained for a region of interest (ROI). This can be the mean or median contrast value within the region, or the value of the contrast at the peak of group activation (Manoach et al., 2001; Wei et al., 2004; Kong et al., 2007; Raemaekers et al., 2007; Friedman et al., 2007). ICCs are then computed for these values. Obtaining one ICC for each activated ROI, one would like to ask if there are significant differences between regional reliabilities. From the ICC inferences in McGraw and Wong (1996), it can be shown that that the low number of subjects, common in neuroimaging experiments, hinders the power to detect ROI differences in ICCs. A typical example can be found in Raemaekers et al. (2007) where they report a highly significant reliability of statistical sensitivity, given by ICC=0.80, $p<0.001$ with confidence interval (0.45,0.94), for a 12-subject experiment. The 95 % confident interval of ICCs is so large that significant differences between ROI reliabilities are difficult to obtain.

A second ICC implementation to compute regional reliabilities is a within-subject measurement (see Raemaekers et al. (2007) for ICC and Specht et al. (2003) for coefficients of determination). Here the reliability of the test–retest signal across ROI voxels is assessed for each subject. This is a measurement of the amount of total variance that can be explained by the intra-voxel variance, and tests the consistency of the spatial distribution of the BOLD signal in a given region, for each individual. Although within-subject ICC is evidently affected by spatial smoothing, it can be used to determine differences between subjects.

A final implementation is the computation of ICC maps (Specht et al., 2003; Aron et al., 2006; Jahng et al., 2005). Although a promising technique, it has not been fully exploited to overcome the limitations of other methods. Aron and colleagues (2006) used voxel-wise ICCs to explore the reliability of activated regions of interest (ROI) for a classification-learning task. They importantly reported the distribution of positive ICC values across a region, and concluded that the relative number of voxels in these regions is higher than in a non-activated area. However they did not examine the whole brain volume or the white matter to account for reliability not associated to the task, nor assigned reliability measures to particular regions.

In the present work we report the reliability of an ROI as the full distribution of ICC values (including negative values) in that region. The reliability distribution is then summarized by its median. This allows us to formally compare the reliabilities across ROIs, increasing the power to detect differences.

The objective of the present study is to explore four aspects of fMRI reliability using the voxel value distributions of ICCs. First, we address the question whether voxels of high group activation in a first scan session are likely to preserve subject differentiability in a second session. In other words, we determine to what degree ICC reliability can be derived from the activation strength of a single session. We therefore evaluate the association between ICC map and the group *t*-map for the first session. The ICC distribution of voxels within the area of high activation is compared with the distribution across the brain and white matter, which are regions not specifically related to the task. This allows us to importantly assess the *relative* increment of the network reliability that can be associated to task response and not, for instance, to non-specific contributors to reliability such as high between-subject variance due to normalization error. Second, we ask whether voxels of high ICC, but low group *t*-value (i.e. not consistently activated) can nonetheless have a consistent behavior across sessions. We consequently select the cluster with highest ICC and suboptimal group *t*-value, and compute the regression of the second-session time-series with that of the first session, for each individual subject. Third, we define the reliability of specific ROIs by the median of their ICC distributions and compare it with three previous implementations, which include the $\mathrm{ICC}_{\mathrm{med}}$ for the ROI medians (Friedman et al., 2007); the $\mathrm{ICC}_{\mathrm{max}}$ at the maximum of group activation (Manoach et al., 2001); and the within measurements (intra-voxel) $\mathrm{ICC}_{\mathrm{v}}$ (Raemaekers et al., 2007; Specht et al., 2003). The comparisons are carried out for different smoothing kernels and cluster sizes, which are assumed to mostly affect $\mathrm{ICC}_{\mathrm{med}}$ and the $\mathrm{ICC}_{\mathrm{v}}$ respectively. Finally, we assess the differences in reliability across activated clusters in order to assess which regions discriminate best between subjects. We have applied these methods to an auditory target detection task, in order to examine simple sensory activations, and an *n*-back task to examine more complex processing in a commonly used paradigm. We chose tasks that activate very different networks to test the robustness of the method.

Materials and methods

Subjects

Ten right-handed, healthy, male volunteers, aged 23–37 (mean 28.7, S.D. 4.6), underwent two scanning sessions separated by three months. Participants were screened for DSM-IV axis I and II disorders using the Structured Clinical Interview for DSM-IV (First et al., 1996). Other exclusion criteria were history of neurological disorders, use of prescription or non-prescription medication that may interfere with interpretation of this study and a score of 8 or more on the Beck Depression Inventory. Participants were asked to refrain from smoking, alcohol and caffeine for a minimum of 48 h before each scanning session. Written, informed consent was provided by each participant for this study, which was approved by the Institute of Psychiatry/South London and Maudsley research ethics committee.

Table 1
Auditory ROIs

ROI	Loc.	*t*-max	Voxels	*t* (sess1 > sess2)
l AC	−56 −6 −4	7.24	190	0.86
r AC	60 −14 2	5.46	94	2.00

Significantly activated clusters (left and right auditory cortex), resulting from the group analysis of 9 subjects for sustained auditory attention greater than base-line. The cluster-corrected significance of $p=0.05$ corresponds to a cluster-defining threshold of $p=0.001$ and cluster size greater than 90 voxels. In the last column we show the mean *t*-score of the ROI for a paired *t*-test between the two sessions, revealing no significant session effects.

To allow familiarization to the equipment and tasks, and minimize practice and learning effects, participants attended a practice session before each scanning session, during which shorter versions of each task were performed.

Auditory target detection

This task required the monitoring of a series of auditory stimuli (a pseudo random sequence of numbers) and identification of targets (number 8) whilst viewing a fixation cross in the centre of a projected computer screen. Numbers were presented via headphones at a rate of 100 per minute for a total of 2 min and 25% of numbers were targets. This block was flanked by control blocks of 30 s that simply required participants to view the fixation cross. Participants were instructed to respond to targets with their right hand using a button box.

N-back

This is a sustained attention task that incorporates a parametric variation in working memory load (Gevins and Cutillo, 1993). Participants were asked to monitor sequentially-presented letters. Each letter was presented for two seconds and participants were asked to press a button with their right hand if the currently presented letter matched the previous letter (‘one back’), two letters before (‘two back’) or three letters before (‘three back’). The control condition required subjects to respond to the letter X (‘look for X’), which occurred at the same frequency as that of the correct targets of the task conditions. Conditions were presented in 42-second epochs and repeated three times in a pseudo-random order. A total of 252 trials were presented over the course of 9 min. Before each epoch an instruction prompt (as indicated above) appeared on the screen for three seconds. The order of the stimuli was pseudo-randomized.

Image acquisition

Participants were scanned on a GE Signa 1.5 T system (General Electric, Milwaukee, WI, USA). A quadrature birdcage coil was utilised for radiofrequency transmission and reception. Thirty-six near axial slices of gradient-echo echoplanar imaging (EPI) data were acquired for each experiment using the following parameters: repetition time (TR = 3000 ms); echo time (TE = 40 ms); flip angle ($\alpha=90°$); slice thickness of 3 mm; interslice gap of 0.3 mm and a 64 × 64 matrix size (3.75 mm × 3.75 mm in-plane resolution). In order to allow the registration of fMRI data to a standard space at a later stage, a higher resolution EPI dataset comprising 43 near axial slices was also acquired with the following parameters: TR = 3000 ms; TE = 40 ms; $\alpha=90°$; slice thickness of 3 mm; interslice gap of 0.3 mm and a matrix size of 128 × 128.

Image processing

The functional data was analyzed using the statistical parametric mapping suite SPM5 (http://www.fil.ion.ucl.ac.uk/spm). Time series were initially realigned. Using a head movement threshold of one voxel resulted in one subject being excluded. A more conservative threshold of 1 mm was subsequently adopted after carrying out intra-voxel reliability analysis, see Fig. 5 and Table 5. The complete analysis was repeated, excluding one subject in the auditory paradigm and two subjects in the *n*-back task.

After movement correction, both test and re-test sessions for each subject were co-registered to the same high resolution EPI, from which the parameters necessary to normalize the BOLD time series to a standard space (MNI template) where obtained. Due to our interest in assessing the effect of smoothing in reliability calculations, unsmoothed time series were fitted to the haemodynamic response model and maps revealing contrasts of interest were obtained for each subject and session. An interim analysis showed that results at group level were similar when smoothing either the time series or the contrast images. The model fitting included movement parameters and a high pass filter of a 160-second period. The contrasts of interest for the auditory and working memory paradigms were those corresponding to sustained attention condition greater than fixation baseline and an average of 1–2–3 back greater than ‘look for x’ baseline.

Intra-subject reliability was calculated at three levels: whole brain, the complete activation network and the activated ROIs. The activation network was obtained using a one sample *t*-test for the first session and a *t*-threshold equivalent to $p=0.001$. Functional ROIs were obtained in a second level analysis from smoothed contrast images at an optimal FWHM = (8 mm, 8 mm, 8 mm) (Smoothing section). The regions were significant at a cluster level ($p<0.05$), corrected for multiple comparisons, and obtained using a voxel-wise threshold $p=0.001$ and 90-voxel extent. The ROIs defined in this way are listed in Tables 1 and 2. For the auditory task bilateral auditory cortex (AC) is activated as expected. For the *n*-back task, all clusters are consistent with regions associated with working memory processes (Smith et al., 1998, Owen et al., 2005), which include the premotor (PMA), and supplementary motor areas (SMA), the dorsolateral prefrontal cortex (DLFPC), the ventrolateral prefrontal cortex (VLPFC), the frontal pole (FP) and the posterior parietal cortex (PPC).

The ROI masks were extracted using the MarsBar tool (Brett et al., 2002). All reliability measures were implemented in dedicated MATLAB toolbox for use with SPM5. The reliability toolbox can be accessed at http://brainmap.co.uk. It performs all implementations presented here, in addition to coefficient of variation analysis for summary statistics across ROIs. It also allows computation of within-subject variance maps that can be used for assessing absolute repeatability of fMRI experiments.

Table 2
N-back ROIs

ROI	Loc.	*t*-max	Voxels	*t* (sess1 > sess2)
l PMA(6)	−32 −4 52	11.21	236	−0.21
r PMA(6)	24 −4 52	10.86	295	0.03
b SMA(6)	−2 12 54	18.19	197	−1.23
l DLPFC(9/45)	−52 12 26	9.68	288	0.13
r DLPFC(9/44)	48 8 24	13.04	218	0.44
l VLPFC(45/47)	−34 24 −2	14.59	180	−0.29
r VLPFC(13/47)	34 20 4	12.32	237	−0.15
r FPC(10)	42 40 26	11.00	366	−0.23
l PPC(7/40)	−42 −44 44	11.64	1131	0.95
r PPC(7)	26 −64 52	12.67	1667	0.70

Significantly activated clusters for the *n*-back task, resulting from the group analysis of 8 subjects. Their relevant location is given with their approximate Brodmann area within brackets (see Image processing section for definitions). The cluster-corrected significance of $p=0.05$ corresponds to a cluster-defining threshold of $p=0.001$ and cluster size of 90 voxels. The regions are consistent with the meta-analysis by Owen et al. (2005). In the last column we show no significant session effect.

Statistical methods

We calculated reliability maps for the third ICC defined by Shrout and Fleiss (1979)

$$\mathrm{ICC}(3,1)=\frac{\mathrm{BMS}-EMS}{\mathrm{BMS}+(k-1)\mathrm{EMS}} \tag{1}$$

Eq. (1) estimates the correlation of the subject signal intensities between sessions, modeled by a two-way ANOVA, with random subject effects and fixed session effects. In this model, the total sum of squares is split into subject (BMS), session (JMS) and error (EMS) sums of squares; and k is the number of repeated sessions.

The maps for the two other ICCs in (Shrout and Fleiss, 1979) showed a similar structure, which confirms the independence of our results from the particular ICC choice. However, the main advantage of Eq. (1) is that it assesses only the level of consistency between measurements (McGraw and Wong, 1996). ICC(3,1) is fully determined by the group effect (F=BMS/EMS). Session effects (F=JMS/EMS) and the group activation for the first session are considered as distinct factors. The latter, the group signal, is estimated by $t=\mathrm{mean}(X_1)\sqrt{(n-1)/\mathrm{std}(X_1)}$; where with X_1 is the first session data, and n is number of subjects.

The assessment of a possible relationship between group activation and reliability was be obtained from the joint probability distribution f(ICC,t). Thus, we computed the distribution of ICCs in the activated region (thresholded by t_0) from

$$f_{t_0}(\mathrm{ICC})=\frac{\sum_{t=t_0}^{\infty} f(\mathrm{ICC},t)}{\sum_{\mathrm{ICC}=-\infty}^{\infty}\sum_{t=t_0}^{\infty} f(\mathrm{ICC},t)} \tag{2}$$

and compared this with the distribution across the whole brain ($t_0=\infty$) and the white matter. The relationship between reliability and activation was tested with a correlation between the threshold (t_0), at which the activated network is defined, and the median of the ICC distribution in the region.

A positive association between ICC and group t does not discard the fact that there can be regions of high ICC but suboptimal t-score. In fact under very specific conditions, group activation of the first session and reliability may increase simultaneously. Such conditions can be derived from the magnitude of group activation (δ) and the correlation (ρ) estimated by the t-statistic and the ICC coefficient, respectively.

The magnitude is given by

$$\delta^2=\frac{n\mu^2}{\sigma_b^2}, \tag{3}$$

where μ and σ_b are the overall group effect and the between-subject variance; while, the correlation coefficient is (McGraw and Wong, 1996)

$$\rho=\frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2} \tag{4}$$

where σ_e is the error variance. A substitution of σ_b of Eq. (4) into Eq. (3) gives

$$\delta^2=\frac{n\mu^2}{\sigma_e^2}\left(\frac{1}{\rho}-1\right) \tag{5}$$

Note that, in particular, for fixed μ and σ_e, the intra-subject reliability actually *decreases* with increasing the level of activation for the first session. This case corresponds to an increment solely in the between-subject variance. A positive association between activation and reliability may result in accompanying reductions in error variance, or increments in overall signal, relative to σ_b. Conversely high reliability can be possible for voxels whose δ and σ_e are both small respect to σ_b. Those are voxels that might fail to fit the model but still have a consistent behavior across subjects.

To assess this possibility we took a 5 mm sphere around the peak voxel within the region of highest ICC that had suboptimal group t. We then extracted the original time series for each subject, averaged across the region, and performed a regression analysis for the second session with the first session to show the stability of the signal.

Spatial correlation of the data has a direct impact upon reliability. Thus, we found it pertinent to assess the effect of smoothing by a given kernel on test–retest reliability, as it is a standard procedure in the analysis of fMRI data. Specifically, we examined the behavior of the median ICC of the brain and task networks with varying smoothing kernels. At this stage we investigated how this pre-processing step can optimize the network and brain reliabilities.

ICC for ROIs

We computed and compared four measures of ROI-ICC. The first was the novel measure computed from ICC maps, while the next three were implementations to neuroimaging data commonly reported in the literature.

- medICC: This is a ROI reliability measure obtained from the median of the ICC distributions within the regions. Its 99% confidence interval can be obtained from the binomial distribution according to (Bland, 2000)

$$j=\mathrm{floor}\left(\frac{n_\nu+1}{2}+2.75\frac{\sqrt{n_\nu}}{2}\right)$$

$$k=\mathrm{floor}\left(\frac{n_\nu+1}{2}-2.75\frac{\sqrt{n_m}}{2}\right)$$

where the limits of the interval are defined by the jth and kth observations of the ordered data, and n_ν is the number of voxels in the region. The median therefore has mean $X_{n\nu/2}$\$ and standard error $(X_j-X_k)/(2*2.75)$. Assuming that the medians are normally distributed, due to their large number of observations, allows us to perform two sample t-tests between ROIs and assess which regions have higher reliability.

- $\mathrm{ICC}_\mathrm{max}$: This is calculated from a summary statistic of the subject activations within the regions. A frequently used statistic is the subject contrast values at the voxel of maximum group activation (Manoach et al., 2001; Kong et al., 2007).
- $\mathrm{ICC}_\mathrm{med}$: In this other implementation ICCs are computed on the median contrast values across the ROIs. Friedman et al. (2007) have utilized the median and show no substantial difference between this and the mean. Since cluster size affects the reliability of the median (Friedman et al., 2007), we calculate ICCs for two cluster sizes. The size of the cluster is indirectly controlled by variations in the voxel-wise threshold. If the t-threshold is incremented then the cluster size is necessarily reduced, since we are looking around regional maxima. We use voxel-wise thresholds of t=4.5 and 7.5. A cluster defined by the threshold t>4.5 has consequently more voxels than at t>7.5. We compare this implementation with medICC. This is done only for the n-back task, since activations on the auditory task include a limited number of voxels above the threshold for statistical significance.

  While $\mathrm{ICC}_\mathrm{max}$ and $\mathrm{ICC}_\mathrm{med}$ can be tested against the null hypothesis (ICC=0) with the F-statistic $F(n-1,\ n-1)$=BMS/EMS (McGraw and Wong, 1996), their confidence intervals are more elaborate functions of F (Shrout and Fleiss, 1979). Note that the intervals for a small sample size are wide, typically including the null value (ICC=0) even for a highly significant estimate.

- ICC$_v$: The last implementation is an intra-voxel measurement (Raemaekers et al., 2007). Here Eq. (1) is applied for each individual subject, using the contrast values of the voxels within the ROIs. ICC$_v$ measures the amount of total variance that can be explained by the intra-voxel variance, and tests the consistency of the spatial distribution of the BOLD signal in a given region, for each individual. Consequently, intra-voxel ICC can be used to determine differences between subjects. In particular we identify changes in signal distribution due to head movement. At a population level we obtain the median and standard deviation from a bootstrapped ICC distribution with a 1000 re-samples of subjects. Since spatial correlation is a key factor for this measure, we examine it prior to smoothing and with an optimal (8 mm) smoothing-kernel and compare it with medICC.

Results

Brain volume and activation network

An ICC brain map was obtained by applying Eq. (1) to the subject contrasts images. The same contrast images, for the first session only, were used to extract a group t-map. Thresholding of t-values corresponding to $p=0.001$ were used to define the activation networks evoked by the tasks. These maps — illustrated in Fig. 1 for the working memory task — show that, although there are some overlapping regions of high ICC and t-values (e.g. parietal cortex), there are other regions of high ICC but low t, and vice-versa. This clearly illustrates the need for a more complete understanding of the relationship between these quantities.

Using the whole brain maps, we calculated an ICC and t-score for each voxel and built the joint probability distribution $f(\mathrm{ICC},t)$, see Fig. 2. The probability distribution of simultaneous high values of reliability and activation appears to depart from the whole brain volume distribution. The distribution for activated voxels was the calculated from Eq. (2) with the corresponding threshold; and the marginal distribution of voxels across the brain was obtained by setting the threshold to $t_0=-\infty$. These distributions are shown in second row of Fig. 2. It is clear that the thresholded voxels tend to have higher values of reliability than voxels across the whole brain volume. Equivalently, the lower right part of the joint distributions are less occupied, which means that there is a low *probability* of finding voxels with high group activation and low reliability. Note that while this is a desired finding, it cannot be formally predicted from general statistical arguments; see Eq. (5).

In addition, we calculated the ICC distribution for the white matter of the brain. This allowed us to compare the distribution of

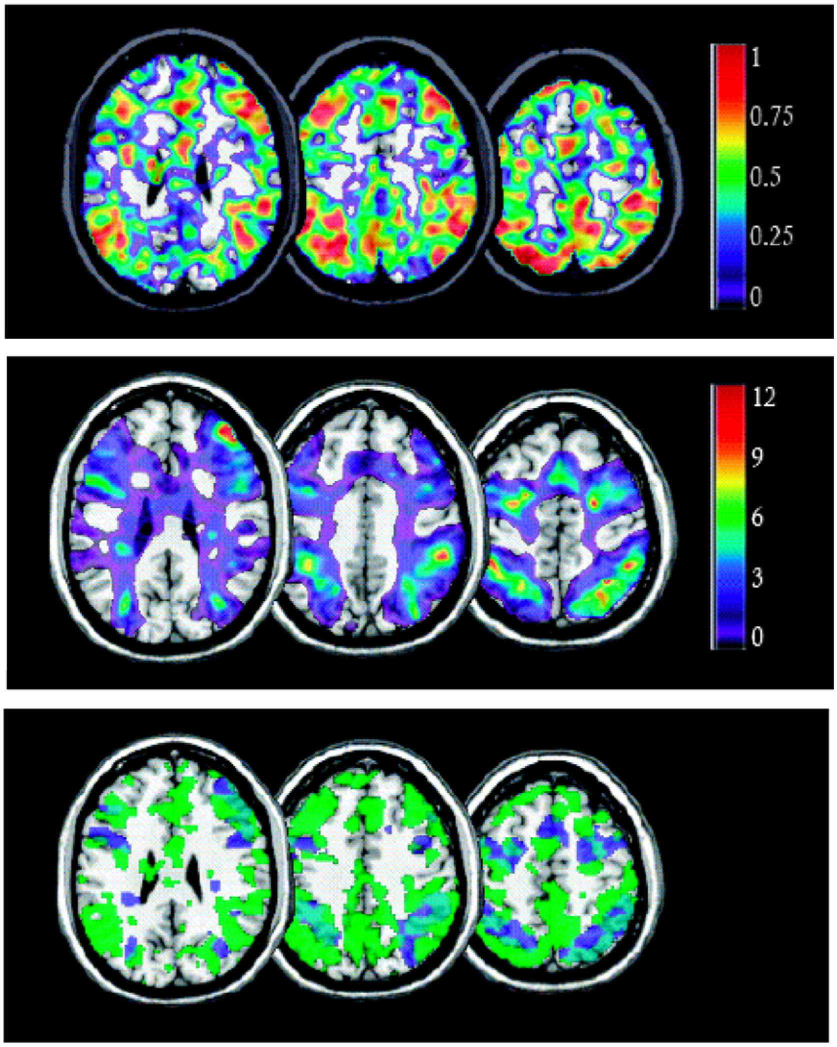

Fig. 1. Top: ICC map, middle: Group t map for first session, bottom: Thresholded regions for the maps above (ICC>0.5 —blue; t>3.5 —green). The figure shows high values ICC(3,1) not necessary follow high values of t. Although there is some overlap (i.e. parietal region), there are large regions of high reliability and low activation. ICC(1,1) and ICC(2,1) (Shrout and Fleiss, 1979) maps are very similar in structure (data not shown).

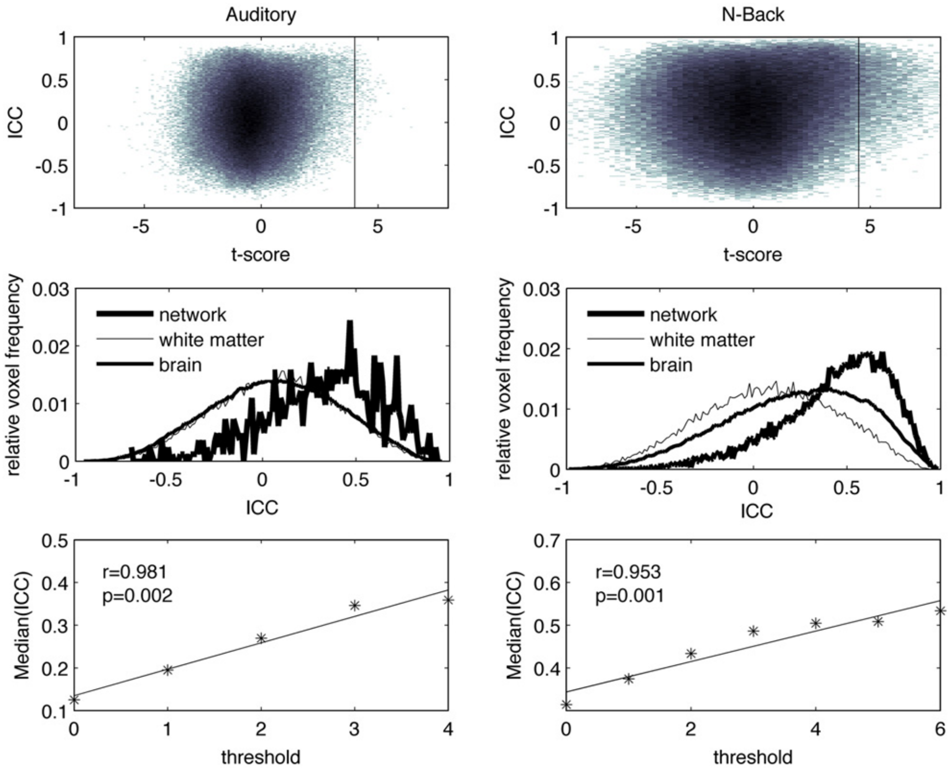

Fig. 2. Joint probability distributions of voxels in the brain for given t and ICC values with appropriate hue to highlight the border effect (top row). The marginal distribution for the whole brain and region of activation are given, for threshold of 4.0 and 4.5 for auditory and n-back tasks (second row). In the case of statistical independence between ICCs and t-scores these two probabilities must be equal. The figure however suggests that high ICCs have more probability of having high t-scores in the thresholded region. This is formally tested on the bottom row, where the median of the network distribution is plotted against the threshold that defines the region. A strong correlation is shown for both tasks. The distribution corresponding to the white matter with negligible median is also shown for both tasks.

the activated regions with a disjoint region that is assumed not to be responsive to the task. For the auditory task we see that the white matter distribution (medICC=0.08, STE=0.006) is similar to that of the brain (medICC=0.07, STE=0.002). The network in this case is mainly composed of the auditory cortex, which is a small region (medICC=0.35, STE=0.03). In the n-back task we see that the whole brain distribution (medICC=0.27, STE=0.002) lies between the white matter (medICC=0.09, STE=0.008) and the activated region (medICC=0.49, STE=0.008) distributions. The task network thus skews the brain volume reliability, while the white matter exhibits a negligible reliability.

In the lower panels of Fig. 2 we show the correlation between the median ICC of the network, defined at different thresholds, and the corresponding threshold values. These strong correlations show that as t_0 is increased the central tendency of the distributions shifts towards more reliable values. Note that a positive relationship between reliability and activation means a direct positive correlation between ICCs and t-scores. Here, however, we have the subtler situation of higher ICCs being found *more frequently* in regions of high activation.

Smoothing

The contrast images were initially calculated without applying any Gaussian smoothing to the data. This allowed us to apply increasing Gaussian kernels to the images, and assess the impact of the smoothing pre-processing on reliability. Fig. 3 shows the ICC medians of the network and whole brain as a function of the FWHM.

Remarkably, we found a consistent optimal value between 8 mm and 10 mm for both regions and tasks. Previous optimal smoothing has been reported between 6.5 mm and 8.5 mm (Worsley, 2005; Gautama and Van Hulle, 2004), using different methods.

With increasing smoothing kernels both BMS and EMS decrease, supporting the simultaneous reduction of intra-subject and error variances. However, it is only their ratio which is optimized by spatial correlation.

High ICC and low t

Fig. 4 displays the correlation of session two with the first session for a peak voxel within a region of high ICC but group t-score below the threshold. We can see that while the time series of the second session fit significantly the first session for five individuals, their t-values for the fit the task model convolved with the HRF are rather low. This translates into a group t (3.2), obtained from contrast values, lower than the network threshold (4.5). Two other subjects presented a low correlation amongst sessions but still consistent low t values. Only for the last subject was the fitting across sessions low and inconsistent from session to session. The consistency in the fitting to the task model, although low, causes the high ICC in and around this region. Note that the 5 mm region used for the computation was close but disjoint from the activated ROI found on the rFPC. Further work on these time-series correlations is currently underway.

ROI reliability

We computed four different reliability measures of the functional ROIs: the median of the ICC distribution in the regions (medICC), the intra-voxel ICC (ICC_v) at a population level, the ICC of the subject contrast medians (ICC_{med}) and the ICC of the contrast values at the maximum of the group activation during the first session (ICC_{max}).

For the auditory cortex we found that although the measures differ substantially, all give a higher reliability on the right hemisphere (see Table 3) compare to the left one. The maximum reliability was obtained

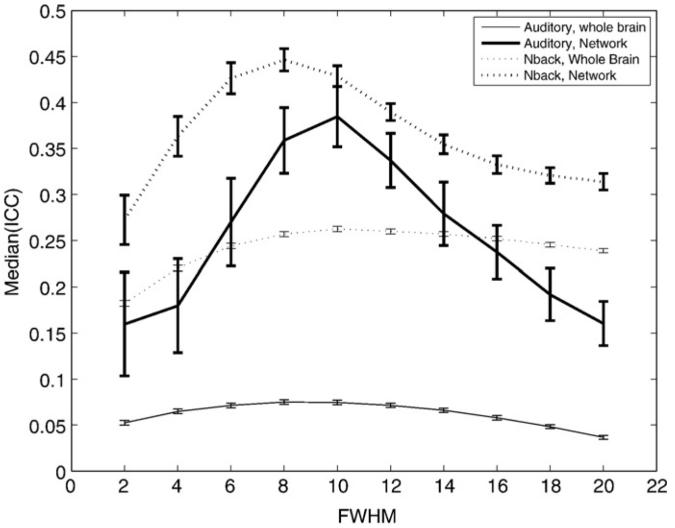

Fig. 3. Median of ICC distribution as a function of the smoothing kernel FWHM in mm. An optimal FWHM is found between 8 mm and10 mm for whole brain volumes and *t*-threshold regions of the two paradigms.

Table 3
Auditory, ROI reliability

ROI	medICC (STE)	ICC_v (STE)	ICC_{med}	ICC_{max}
l AC	0.20 (0.05)	0.52 (0.1)	0.03	−0.10
r AC	0.55 (0.06)	0.56 (0.2)	0.57	0.35

Four reliability measures for the functional ROIs of the auditory task: the median ICC in the ROI, the intra-voxel reliability, the ICC of the median of the contrast values, and the ICC of the contrast values at the voxel with maximum group activation. All consistently show the left auditory cortex as less reliable.

with the ICC_v. For the right cortex, medICC and ICC_{med} reported similar values. The left cortex had a negligible reliability for the latter measure. The ICC_{max}, based on the single most activated voxel, reported the lowest reliability for both ROIs. ICC_{med} and ICC_{max} are correlation measurements that can be directly assessed against the null hypothesis with an *F* test with eight degrees of freedom ($F(8,8)=BMS/EMS$). The critical ICC corresponding to a $p=0.05$ is $ICC=0.54$ with 95% confidence interval (−0.11, 0.87). Therefore for the auditory cortex, only ICC_{med} in the right auditory cortex was significant.

While ICC calculated from summary are expected to be more variable, due to the low number of subjects typical in neuroimaging experiments, ICC_v reported the highest values as consequence of spatial autocorrelation. These general observations were also found for the *n*-back task see Table 4, the highest and most relevant highlighted in bold face.

As with the auditory, task we observed a degree of consistency in terms of the ICC ranking of ROIs across all measures, despite of absolute reliability figures being substantially different. Because the *n*-back task offered more ROIs, this consistency was explored in more detailed by correlating all ICC implementations across ROIs and under variations of smoothing and cluster size (Table 4). Here we show the reliabilities of all the ROIs ordered according to medICC for smoothed data and cluster size obtained with a *t*-threshold (t>4.5); the left most column. Correlations between ROI reliabilities obtained by different methods are shown at the bottom of the table.

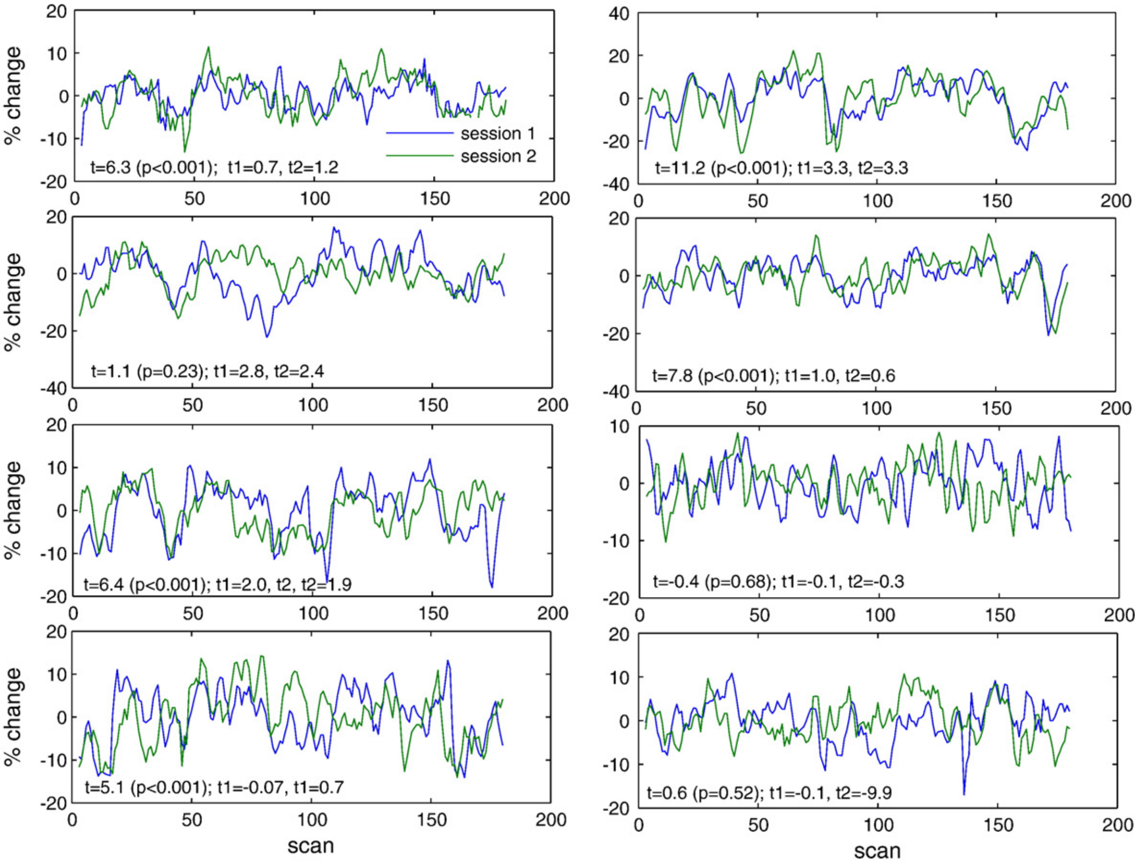

Fig. 4. Time series of percent signal change from global mean within a 5 mm sphere around MNI coordinates (46 32 34) corresponding to a high ICC (0.97) but group *t*-score (3.2) below statistical threshold (4.5). *T*-scores at the bottom of each plot are the product of the regression of the second session with the first session (t), and the regression of the first ($t1$) and second session ($t2$) to the HRF, given by the values of spm *t* maps at that location. A linear trend has been removed form the original time series to account for low frequency fluctuations.

Table 4
N-back, ROI reliability

	medICC(STE)			ICC$_v$(STE)		ICC$_{med}$		ICC$_{max}$	σ_b^2
FWHM	8 mm	0 mm	8 mm	8 mm	0 mm	8 mm	8 mm	8 mm	
Threshold	t>4.5	t>4.5	t>7.5	t>4.5	t>4.5	t>4.5	t>7.5	NA	
rFPC	0.64 (0.03)	0.42 (0.02)	0.64 (0.03)	0.56 (0.08)	0.43 (0.06)	0.75	0.77	0.59	0.24
rPPC	0.56 (0.01)	0.50 (0.01)	0.53 (0.05)	0.74 (0.70)	0.56 (0.08)	0.28	0.36	0.27	0.41
lPPC	0.53 (0.01)	0.47 (0.01)	0.63 (0.03)	0.69 (0.08)	0.53 (0.08)	0.30	0.55	0.36	0.34
rDLPFC	0.48 (0.02)	0.51 (0.02)	0.51 (0.05)	0.71 (0.06)	0.47 (0.12)	0.28	0.44	0.36	0.34
rVLPFC	0.46 (0.03)	0.27 (0.03)	0.38 (0.06)	0.50 (0.14)	0.29 (0.06)	0.47	0.39	0.39	0.21
lDLPFC	0.36 (0.02)	0.36 (0.02)	0.01 (0.09)	0.62 (0.09)	0.47 (0.06)	−0.40	−0.15	0.01	0.12
lPMA	0.36 (0.04)	0.36 (0.02)	0.53 (0.06)	0.62 (0.09)	0.47 (0.06)	−0.41	0.39	0.61	0.12
rPMA	0.29 (0.02)	0.30 (0.02)	0.15 (0.04)	0.65 (0.06)	0.43 (0.03)	0.19	0.07	0.15	0.16
bSMA	0.27 (0.04)	0.28 (0.04)	0.14 (0.11)	0.65 (0.07)	0.45 (0.03)	−0.04	−0.09	−0.16	0.21
lVLPFC	−0.01 (0.04)	0.09 (0.04)	−0.12 (0.08)	0.41 (0.06)	0.25 (0.05)	−0.11	−0.23	−0.12	0.008
A									
(r,p)	(1,0)	(0.86,0.001)	(0.86,0.001)	(0.55,0.09)	(0.60,0.06)	(0.62,0.05)	(0.85,0.001)	(0.72,0.02)	(0.79,0.005)
(r,p)		(1,0)	(0.79,0.006)	(0.83,0.002)	(0.85,0.001)	(0.33,0.34)	(0.68,0.02)	(0.56,0.08)	
(r,p)			(1,0)	(0.48,0.15)	(0.53,0.01)	(0.55,0.09)	(0.96,10^{-4})	(0.87,0.001)	
B									
(r,p)				(1,0)	(0.93,10^{-4})	(0.03,0.91)	(0.29,0.4)	(0.18,0.61)	
(r,p)						(1, 0)	(0.69,0.03)	(0.42,0.15)	
C									
(r,p)							(1,0)	(0.90,10^{-4})	

Four reliability measures for the functional ROIs of the *n*-back task: the median ICC in the ROI (medICC), the intra-voxel reliability (ICC$_v$), the ICC of the median of the contrast values (ICC$_{med}$), and the ICC of the contrast values at the voxel with maximum group activation (ICC$_{max}$). Correlation coefficients between the columns are shown in the bottom rows. The reference column is indicated as (1,0). Section A shows the correlation of medICC with the other measures; section B gives the correlations within ICC$_v$ and ICC$_{med}$; and the correlation between ICC$_{med}$ and ICC$_{max}$ is illustrated in section C. Last column represents the median (squared) between-subject variance in each ROI.

In Table 4, we also include the median of the ICC$_v$ and ICC$_{med}$ for prior-to-smooth data and reduced cluster size data obtained, respectively. We compared these measures with medICC under those conditions.

The correlations are classified in three different categories. The first set of values (A-rows) corresponds to the comparison of medICC with all other measures. From the first row, we can see that the medICC is robust under changes in smoothing kernel and cluster size. Furthermore the magnitudes of the measures are kept within a common range. Note that the first column provides the highest and most significant correlations across all other columns.

In the second and third column of section A, we show the correlations between medICC and ICC$_v$, and ICC$_{med}$ for prior-to-smooth data and the reduced cluster size, respectively. The strong correlations suggest that ICC$_v$ and ICC$_{med}$ become equivalent to medICC under the above conditions.

Section B shows the relative impact of smoothing by a Gaussian kernel and cluster size on ICC$_v$ and ICC$_{med}$. The first row displays the correlation of ICC$_v$ between prior-to-smooth data and smoothed data with an 8 mm kernel. We observe that although ICC$_v$ is highly susceptible to spatial correlation, the rank order of ROIs reliabilities is preserved, shown by a high correlation between smoothing conditions. Note that ICC$_v$ correlates best to medICC for unsmoothed data, while it has non-significant correlations with ICC$_{med}$ and ICC$_{max}$.

In the second row of section B, ICC$_{med}$ exhibits poor correlation for two different cluster sizes, corresponding to different *t*-thresholds. Friedman et al. (2007) already acknowledged cluster instability of the ICC calculated from the median activated voxel. In particular, Friedman and colleagues 2007 argue in favor of increasing cluster size in the calculation of ICCs for median values. Our analysis, on the other hand, supports the view that reduced cluster size leads to better correlation with other ICC measures like medICC and ICC$_{max}$, shown in section C of Table 4. This issue is, nevertheless, overshadowed by the fact that both ICC$_{med}$ and ICC$_{max}$ have larger variance across the ROIs. The confidence intervals for the critical ICC value can, again, explain this. For the *n*-back task a significant (p=0.05) ICC is reached at 0.58 [F(7,7)=BMS/EMS] with a corresponding 95% interval of (−0.10, 0.89). This is rather large, allowing for high ICC variance. With this critical ICC we see, for instance, that only the rFPC is significantly reliable for these two measures, and the power to detect ICC differences is greatly reduced. This is not the case with medICC, which we can be used to determine reliability differences across ROIs.

The values of the first column of Table 4 are illustrated in Fig. 5, from which we can determine the ROIs that discriminate best amongst the individuals of the sample. Differences in reliability between ROIs were formally evaluated with a multiple comparison test (two sample *t*-test), where the medians are assumed to distribute normally. The table shows that the activated network is more reliable than the white matter and whole brain volume. Of the clusters belonging to the network, rFPC and the parietal cortex presented the highest reliability. Whereas, the pre-motor and the supplementary motor area were less reliable. It is noteworthy that the lVLPFC has null reliability. The reliabilities of the rVLPFC and the DLPFC fell between the reliabilities for the brain and network.

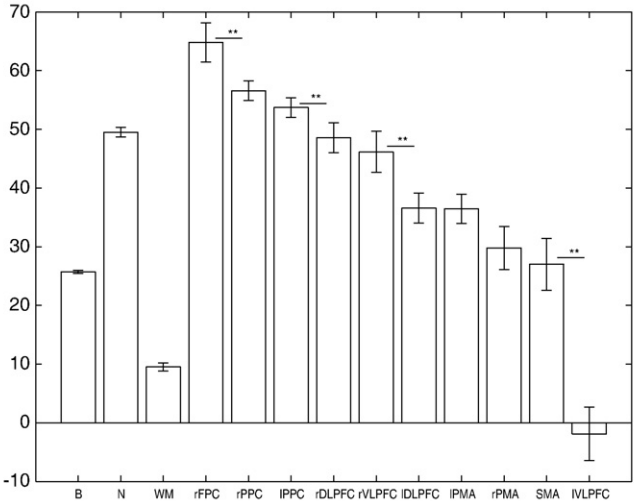

Fig. 5. Medians of the ICC distributions in each ROI for the *n*-back task. The figure shows also the reliabilities for the brain (B), network (N) and white matter (WM). A multiple comparison test was run across all the ROIs (α=0.01/45). Significant differences are shown.

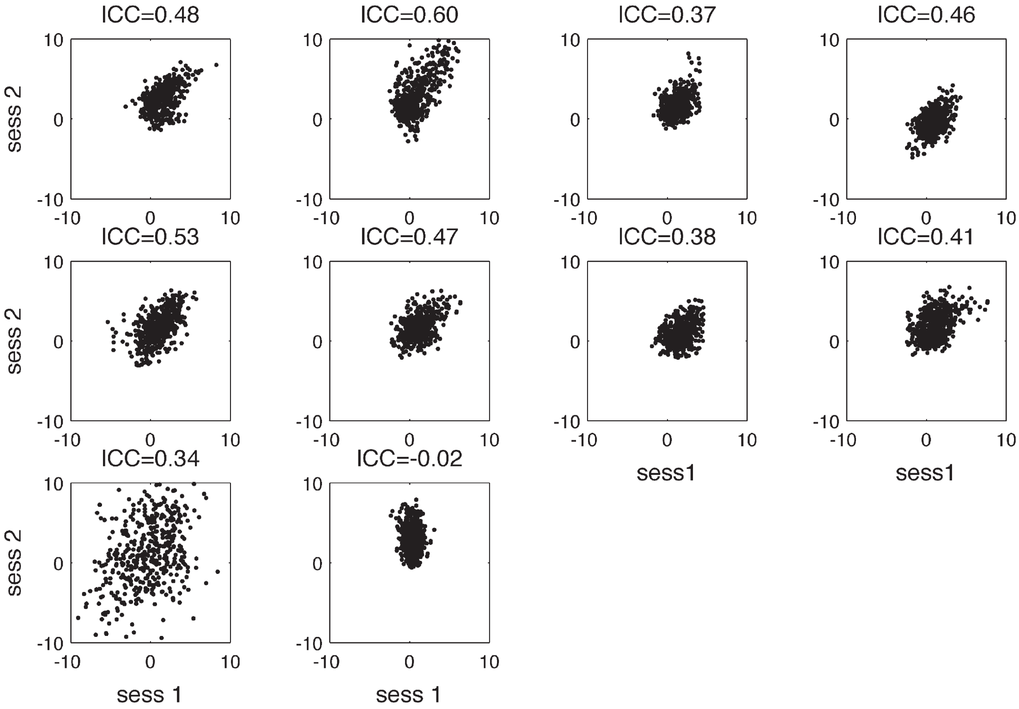

Fig. 6. Comparison between session 1 and session 2 for each subject. The figures show the contrast values for each voxel within the SMA. On top of each figure, the measures of voxel reliability are given. The two last plots reveal the effect of movement on reliability. These two subjects were excluded from the analysis for having movement greater than 1 mm. It is notable that the subject on the left moved considerably on both sessions, whereas the subject on the right moved mostly on the second session.

Movement

In this study we excluded one subject from the analysis of both tasks, and an additional one from the *n*-back task. Using $\mathrm{ICC_v}$ for non-smoothed data, we explored the reliabilities of each subject individually. Note that this is not possible with the other ICC implementations. Fig. 6 shows the calculation of the intra-voxel ICC for all ten subjects within the SMA, which was the region of strongest activation at peak voxel. The figure shows the contrast values for each voxel in session 1 and session 2, for the *n*-back task. The last two plots correspond to the subjects discarded by their high head movements who were outliers in terms of $\mathrm{ICC_v}$ averaged across all regions, see Table 5.

We found that head movement had also great impact on medICC, as can be predicted from its strong correlation with $\mathrm{ICC_v}$ for unsmoothed data. Indeed, the inclusion of the two volunteers for the *n*-back task leads to a substantial reduction in medICC across the whole brain (medICC=−0.04, STE=0.002), network (medICC=0.08, STE=0.009) and white matter (medICC=−0.24, STE=0.006). In this way, outliers in terms of movement can be also identified using medICC.

Discussion

In this paper we have presented a robust implementation of test–retest reliability based on the intra-class correlation coefficient. The method extends the voxel-wise calculation of ICCs, based on the medians of ICC *distributions* of given regions. This measure allows the assessment of the reliability of the activated network relative to the whole brain volume and white matter. It also enables comparisons of the reliabilities across regions of activation, revealing which activated cluster discriminate best amongst individuals.

We found that although there is not a formal relationship between the first session *t*-map and the ICC map, there is a level of association between them. Voxels with high group-*t* had higher probability of high ICC. This enables us to say that voxels that are found active on a single session have greater chance to show consistent activation across subjects if subsequent sessions are performed. Since the ICC can be regarded as a measure of consistent differentiability between subjects, we can also say that the areas activated during a first scan session are more *likely* discriminate between subjects than other areas not engaged with the task.

Raemaekers and colleagues (2007) have shown that the width of the distribution of *t*-values across the brain accounts for much of the reliability of activated areas. Similarly, we show that for the *n*-back task the distribution of ICCs across the *whole* brain is skewed towards positive values. Both findings show an intrinsic reliability that should be associated to more general conditions of the experiment. Note that such conditions were different for the auditory task since for this case the whole brain distribution was not skewed. Interestingly, the white matter distribution, which by principle is not related to the task, has negligible reliability for both cases. Importantly, for both these tasks reliability of the network, or specific regions can be assessed against these measures of intrinsic reliability.

The general association between the ICC and *t*-maps can be regarded as a feature of fMRI data, since it is not possible to formally derive ICC scores from group *t*-values of the test session, see Eq. (5). The interesting possibility that voxels with high ICC but low *t* is then opened. In fact, there are voxels whose relation to the task is not directly measured by the fitting of their task model (convolved with the HRF).

For the *n*-back task a cluster near the activated region in the rFPC had very high reliability but suboptimal *t*. Importantly, we found that

Table 5
Outliers due to movement (*n*-back)

Subject	1	2	3	4	5	6	7	8	9	10
<ICCv>[a]	0.43	0.59	0.54	0.31	0.40	0.41	0.37	0.38	0.02	0.06
Max. trans. mm	0.42	0.33	0.22	0.43	0.71	0.16	0.24	0.55	4.32	1.35

The reliabilities for the two last subjects fall around 3 standard deviations away from the mean of the first eight subjects.

[a] ICCs are averaged across all ROIs of each subject.

the time series for most subjects highly correlate across sessions and consistently fail to fit the task model significantly. The existence of such voxels suggests that there are regions that respond indirectly, or non-linearly, to the stimuli, yet they can appear in meaningful regions associated to relevant brain function. A deeper investigation on these issues is beyond the scope of the current analysis, but is clearly important.

We have used voxel-wise ICCs for the general purposes of measuring the reliability of activated ROIs, showing its robustness in relation to other ICC implementations for neuroimaging data, and demonstrating its potential suitability to identify head motion and optimal smoothing kernels.

We have found that medICC is less variable across ROIs than the measures based on summary statistic, $\mathrm{ICC}_{\mathrm{med}}$ and $\mathrm{ICC}_{\mathrm{max}}$. These former measures are particularly sensitive to small number of observations. The 95% confidence interval for the critical ICC value, for typical number of subjects in neuroimaging experiments, is very wide, allowing for large errors produced by poor estimations. This is the most likely reason for their numerical discrepancy with the other ICC implementations.

More importantly, we show a consistent ranking of the ROI reliability across all the ICC implementations. The correlations of medICC with $\mathrm{ICC}_{\mathrm{v}}$, for non-smoothed data, and with $\mathrm{ICC}_{\mathrm{med}}$, for reduced cluster size, were both strong. The correlation with $\mathrm{ICC}_{\mathrm{max}}$ was stronger when the cluster size was reduced. In these terms, the medICC can account for other ICC implementations. It is important to note that medICC at optimal smoothing kernel and lower *t*-threshold (4.5) gave the best correlations compared with all other ICCs. Others measures did not perform as good; in particular, $\mathrm{ICC}_{\mathrm{v}}$ did not correlate strongly with other measures.

From these comparisons, we have also found that medICC was more robust under conditions of different statistical thresholding and Gaussian smoothing. Specifically, the correlations of medICC between smoothed and prior-to-smooth data, and between two cluster sizes, were particularly high. On the contrary, $\mathrm{ICC}_{\mathrm{med}}$ and $\mathrm{ICC}_{\mathrm{v}}$ were greatly affected by the level of the threshold and the spatial correlation, respectively.

A plausible reason to still prefer $\mathrm{ICC}_{\mathrm{med}}$ and $\mathrm{ICC}_{\mathrm{max}}$ to medICC is to account for spatial normalization error. However, it should not be assumed that voxel-wise ICCs require perfect normalization. The evaluation of an ICC for a given voxel assumes that the observation on that region is *equivalent* to a second observation around the same coordinates. In this sense an ICC evaluated on the voxel of maximum group activation can also be justified. The equivalence between the voxels at the same coordinates is supported by the idea that the parenchyma within that region, for each session, if not identical, is at least functionally equivalent. If normalization (or matching between parenchyma) is improved we should expect to have a closer estimate of the true ICC for that particular voxel. In this case voxel-wise ICCs could be even used to assess normalization error.

As for ICCv, we confirmed that its overall level depends strongly on smoothing. This is a consequence of being obtained from non-independent observations. A direct interpretation in terms of intra-class correlation is not as yet possible, since spatial correlation affects the level of heterogeneity of the data. We have shown, however, the usefulness of the measure in two different aspects. First, the ranking of the ROIs is highly consistent with medICC, independent of the level of smoothing. Spatial correlation has a major impact on level of $\mathrm{ICC}_{\mathrm{v}}$ but not on the ROI ranking as given by medICC. This shows again the importance of the *relative* classification of reliability across regions. Secondly, the measure can be used at an individual level. Its application to determine movement outliers illustrates how subjects with unlikely reliability can be identified.

The measure introduced here, medICC, opens the possibility of formally comparing the reliability of defined regions or comparing the same ROIs across different groups dependent on the differential heterogeneity of the samples (see below). The measure is based on *all* the observations for a region, improving the detection of reliability differences amongst ROIs. The confidence intervals of our reliability calculation are tight compared with other methods. This allowed us to formally compare which of the regions can discriminate best between individuals. We found that although most of the ROIs *tended* to have higher reliability than the whole brain, there were regions of high activation and low reliability. This suggests that reliable activated ROIs might have a different functional role than unreliable, yet activated, regions.

Unfortunately this study was not powered sufficiently to test correlations of performance measures and brain activation across the sessions. However, using the functional definitions of Owen et al. (2005) for the *n*-back task, we see for instance that the control of several cognitive processes (rFPC) is the most reliable, together with the more fundamental processes of short-term storage (l–r PPC). Therefore, with respect to these processes, subjects are more differentiable. More automatic responses like planning motor action (SMA), visuospatial attention and readiness to respond (lPMA) are more uniform across subjects, despite their high level of activation. Strategic reorganization and control (DLPFC), on the other hand, can be considered reliable. Interestingly the region of lowest reliability was the VLPFC, which is thought not to have an evident role in working memory (Rushworth et al., 1997), but may be more involved in response mapping (Owen et al., 2005).

It is worth remembering that the ICC depends on both the within and the between-subject variances. Although changes in between-subject variance can drive changes in the ICC, as now shown in Table 4, this variance alone does not fully provide subject discrimination. While it would be desirable to have large differences between subjects, high discrimination cannot be established without simultaneously having a relative small within-subject variance. This guarantees the repeatability of the measures and, therefore, the proper assignment of observations to subjects and assessment of individual tracking.

A consequence of the strong correlation of the ICC with the between-subject variance across ROIs is that changes in repeatability (reliability associated to error measurement) are not adequately represented with changes in ICC. The coefficient is clearly dependent in the heterogeneity of the sample. Although this is a convenient feature for assessing differences in subject heterogeneity (see Manoach et al., 2001), the ICC cannot be generalized as an assessment of measurement error.

This is a clear limitation of the method since a fundamental part in the development of fMRI is to determine the error measurements that can establish which observations are clinically acceptable. Due to the wide range of scanning protocols, task and analysis methods this challenge is however far from simple. On this direction, Zandbelt et al. (2008) use the within subject variance to directly assess the error measurement produced when estimating the “true” average of each subject's activation. The relative units of BOLD based fMRI pose an additional challenge. Given that units depend on the subject's baseline it is unclear how to identify a clinical standard for the measurement error. Measurements of reliability and repeatability on physiological units can be more suitable for this purpose.

In the present study a sample size of ten was used throughout the development of the methodology. While this can be considered a small sample size, we have used two different tasks to substantiate the results. In addition, we have tested the robustness of the method using an anti-saccade task involving 25 subjects and found similar results. This is the subject of a separate manuscript.

In conclusion, we have introduced a robust measure of regional reliability, based on the voxel-wise calculation of ICCs. This ICC implementation can be used to formally compare reliability across brain regions. More generally, the measure shows *greater* intra-subject reliability in the active network than that of the whole brain volume or regions not engaged by the task. In these *relative* terms, the brain activation of the tasks studied can be considered reliable.

Acknowledgments

We thank Professor Michael Brammer for useful discussion and GlaxoSmithKline for financial support in the collection of the data.

References

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage 29 (3), 1000–1006.

Bland, J., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1 (8476), 307–310.

Bland, M., 2000. An Introduction to Medical Statistics. Oxford University.Press, Oxford, UK.

Bland, J., Altman, D., 1996. Statistics notes: measurement error and correlation coefficients. BMJ (313), 41–42.

Brett, M., Anton, J., Valabregue, R., Poline, J., 2002. Region of interest analysis using an SPM toolbox. NeuroImage 16 (2), 2–6.

First, M., Spitzer, R., Gibbon, M., Williams, J., 1996. Structured Clinical Interview for Axis I Disorders—Patient Edition. New York Biometrics Research, New York State Psychiatric Institute, New York.

Friedman, L., Stern, H., Brown, G., Mathalon, D., Turner, J., Glover, G., Gollub, R., Lauriello, J., Lim, K., Cannon, T., Greve, D., Bockholt, H., Belger, A., Mueller, B., Doty, M., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S., 2007. Test–retest and between site reliability in a multicenter fMRI study. Hum. Brain Mapp.

Gautama, T., Van Hulle, M.M., 2004. Optimal spatial regularisation of autocorrelation estimates in fMRI analysis. Neuroimage 23 (3), 1203–1216.

Gevins, A., Cutillo, B., 1993. Spatiotemporal dynamics of component processes in human working memory. Electroencephalogr. Clin. Neurophysiol. 87 (3), 128–143.

Jahng, G. -H., Song, E., Zhu, X. -P., Matson, G.B., Weiner, M.W., Schuff, N., 2005. Human brain: reliability and reproducibility of pulsed arterial spinlabeling perfusion MR imaging. Radiology 234 (3), 909–916.

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

Kiehl, K.A., Liddle, P.F., 2003. Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test–retest study. Hum. Brain Mapp. 18 (1), 42–52.

Kong, J., Gollub, R.L., Webb, J.M., Kong, J. -T., Vangel, M.G., Kwong, K., 2007. Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. Neuroimage 34 (3), 1171–1181.

Liu, J.Z., Zhang, L., Brown, R.W., Yue, G.H., 2004. Reproducibility of fMRI at 1.5 t in a strictly controlled motor task. Magn. Reson. Med. 52 (4), 751–760.

Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y., Goff, D.C., Rauch, S.L., Kennedy, D.N., Gollub, R.L., 2001. Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158 (6), 955–958.

McGraw, K., Wong, S., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1 (1), 30–46.

Owen, A., McMillan, K., Laird, A., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum. Brain Mapp. 25 (1), 46–59.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36 (3), 532–542.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Valk, J., Scheltens, P., 1997. Test–retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am. J. Neuroradiol. 18 (7), 1317–1322.

Rushworth, M.F., Nixon, P.D., Eacott, M.J., Passingham, R.E., 1997. Ventral prefrontal cortex is not essential for working memory. J. Neurosci. 17 (12), 4829–4838.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.

Smith, E.E., Jonides, J., Marshuetz, C., Koeppe, R.A., 1998. Components of verbal working memory: evidence from neuroimaging. Proc. Natl. Acad. Sci. U. S. A. 95 (3), 876–882.

Specht, K., Willmes, K., Shah, N.J., Jancke, L., 2003. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 17 (4), 463–471.

Tjandra, T., Brooks, J.C.W., Figueiredo, P., Wise, R., Matthews, P.M., Tracey, I., 2005. Quantitative assessment of the reproducibility of functional activation measured with BOLD and MR perfusion imaging: implications for clinical trial design. Neuroimage 27 (2), 393–401.

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

Worsley, K.J., 2005. Spatial smoothing of autocorrelations to control the degrees of freedom in fMRI analysis. Neuroimage 26 (2), 635–641.

Yoo, S. -S., Wei, X., Dickey, C.C., Guttmann, C.R.G., Panych, L.P., 2005. Long-term reproducibility analysis of fMRI using hand motor task. Int. J. Neurosci. 115 (1), 55–77.

Zandbelt, B.B., Gladwin, T.E., Raemaekers, M., van Buuren, M., Neggers, S.F., Kahn, R.S., Ramsey, N.F., Vink, M., 2008. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. Neuroimage 42 (1), 196–206.