ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 36 (2007) 532–542

Test–retest reliability of fMRI activation during prosaccades and antisaccades

M. Raemaekers,[a,b,*] M. Vink,[b] B. Zandbelt,[b] R.J.A. van Wezel,[a] R.S. Kahn,[b] and N.F. Ramsey[b]

[a]*Helmholtz Institute, Department of Functional Neurobiology, University of Utrecht, The Netherlands*
[b]*Rudolf Magnus Institute of Neuroscience, Department of Psychiatry, University Medical Center Utrecht, The Netherlands*

Received 27 September 2006; revised 14 February 2007; accepted 13 March 2007
Available online 11 April 2007

Various studies have investigated reproducibility of fMRI results. Whereas group results can be highly reproducible, individual activity maps tend to vary across sessions. Individual reliability is of importance for the application of fMRI in endophenotype research, where brain activity is linked to genetic polymorphisms. In this study, the test–retest reliability of activation maps during the antisaccade paradigm was assessed for individual and group results. Functional MRI images were acquired during two sessions of prosaccades and antisaccades in twelve healthy subjects using an event-related fMRI design. Reliability was assessed for both individual and group-wise results. In addition, the reliability of differences between subjects was established in predefined regions of interest. The reliability of group activation maps was high for prosaccades and antisaccades, but only moderate for antisaccades vs. prosaccades, probably as a result of low statistical power of individual results. Reproducibility of individual subject maps was highly variable, indicating that reliable results can be obtained in some but not all subjects. Reliability of individual activity maps was largely explained by individual differences in the global temporal signal to noise ratio (SNR). As the global SNR was stable over sessions, it explained a large portion of the differences between subjects in regional brain activation. A low SNR in some subjects may be dealt with either by improving the statistical sensitivity of the fMRI procedure or by subject exclusion. Differences in the global SNR between subjects should be addressed before using regional brain activation as phenotype in genetic studies.

Keywords: fMRI; Reliability; Antisaccades; Intraclass correlation

Introduction

Functional imaging is becoming increasingly popular for linking differences in information processing to genetic polymorphisms (Hariri and Weinberger, 2003). Some studies have shown that fMRI measurements can be more sensitive for revealing abnormalities than the corresponding behavioral measures such as task performance (Callicott et al., 2003; Raemaekers et al., 2006a; Vink et al., 2006). The penetrance of the genes underlying the neural mechanisms may thus be higher when the expression is measured at the neurofunctional level than when measured at the behavioral level. This suggests that fMRI images could allow for more accurate phenotyping than behavioral measures, and are therefore better endophenotypes. However, one of the prerequisites of a good endophenotype is that it is a heritable and trait-like characteristic (Gottesman and Gould, 2003), and should thus be state-independent, and have a high test–retest reliability. In this study, we assess the reliability of brain activation maps that are associated with a well-known endophenotype for schizophrenia: the antisaccade paradigm.

In the antisaccade task, subjects have to inhibit an eye movement towards a novel stimulus (prosaccade) and instead make an eye movement in the opposite direction (antisaccade) (Hallett, 1978). Patients with schizophrenia (Fukushima et al., 1990) and their relatives (Calkins et al., 2004) have difficulty suppressing reflexive saccades during the antisaccade task. The antisaccade paradigm may thus become a useful tool for identifying those with a genetic risk for schizophrenia (Hutton and Ettinger, 2006). Impaired antisaccade performance has also been observed in healthy co-twins of schizophrenic patients, providing more direct evidence for a link of antisaccade performance to genetic liability for schizophrenia (Ettinger et al., 2006). The test–retest reliability of reaction times of prosaccades and antisaccades ranges from fair to good across studies in healthy subjects (Ettinger et al., 2003; Harris et al., 2006; Klein and Berg, 2001; Roy-Byrne et al., 1995) and also in patients with schizophrenia and their unaffected relatives (Calkins et al., 2003; Harris et al., 2006). Most, but not all of these studies report good reliability for antisaccade error rates as well.

* Corresponding author. Department of Functional Neurobiology, Helmholtz Institute, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Fax: +31 30 2505443.
E-mail address: m.a.h.l.l.raemaekers@bio.uu.nl (M. Raemaekers).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.03.061

The assessment of test–retest reliability in fMRI studies is a more complicated matter. Similar to behavioral measures, reproducibility of fMRI is subject to cognitive factors such as arousal, cognitive strategies, and learning. In addition to actual differences in brain function between sessions, fMRI measures are also influenced by various non-psychological factors such as changes in the position of the subject in the magnetic field of the MRI scanner and in the radiofrequency head coil, field inhomogeneities, image signal to noise ratio (SNR), and cardiac, respiratory, and motion artifacts (McGonigle et al., 2000; Veltman et al., 2000). All these factors can seriously affect image reproducibility.

Several measures are available to assess the reliability of fMRI data. For estimating the reliability of individual and group-wise results, one can look at the intraclass correlation of contrast *t*-values for pairs of activation maps (ICC_{within}) (Shrout and Fleiss, 1979), or at the ratio of overlapping activation ($R^{12}_{overlap}$) (Machielsen et al., 2000; Rombouts et al., 1998). For using fMRI images as endophenotype, the reliability of differences between subjects is the critical factor. The between-subject reliability is reflected by the intraclass correlation across subjects over repeated sessions ($ICC_{between}$) (Shrout and Fleiss, 1979). The $ICC_{between}$ is calculated by taking the ratio of the variance between subjects and the variance within subjects over the repeated measurements. A high $ICC_{between}$ reflects a large between-subject variability, and a small within-subject variability. A good endophenotype therefore needs a high $ICC_{between}$.

A limited number of studies have assessed the $ICC_{between}$ of fMRI task activation. Poor reproducibility was found for a verbal working memory task which was repeated 9 times with an interval of 3 weeks (Wei et al., 2004). In contrast, Manoach et al. (2001) found moderate reliability in healthy subjects, also during a working memory task, and a recent study of Aron et al. (2006) reported good to nearly perfect reproducibility in 8 subjects in a 1-year follow-up study for a classification learning task, especially in the frontostriatal circuitry. The latter finding suggests that fMRI could already be used for phenotyping based on patterns of regional brain activation.

These studies did not report estimates of the reliability for results of *individual* subjects, such as the ICC_{within} or the $R^{12}_{overlap}$. Although this seems trivial, as normally a high $ICC_{between}$ cannot exist in the absence of reliable within-subject measurements, the case may be different for fMRI. The first prerequisite for finding reproducible individual results in fMRI is a good statistical sensitivity, and thus a good temporal SNR (technically contrast to noise ratio, CNR). Noisy data will by definition be less reproducible. Subjects may greatly differ in their global temporal SNR, and therefore in their extent of total activation. Such difference could have no relationship to actual differences in brain activation, but merely reflect differences in, e.g., global physiological noise, motion induced noise, or properties of the hemodynamic response. Differences between subjects in the global SNR could very well be stable over time. A high $ICC_{between}$ in a voxel or a region of interest (ROI) could therefore reflect stable individual differences in the SNR across the whole brain, instead of regionally specific differences. For $ICC_{between}$ estimation, it is therefore important to take the global temporal SNR into account.

In this study we address the reproducibility of brain activation maps during prosaccades and antisaccades, and the contrast between prosaccades and antisaccades at the group level as well, as at the individual level. This is done by calculating ratios of overlapping activation and intraclass correlations of group-wise *t*-maps and *t*-maps of individual subjects. Group results of two sessions will be compared to detect systematic changes in brain activation between the sessions. The $ICC_{between}$ will be calculated for brain activation in predefined ROIs. In addition, we propose a method for assessing the global SNR in individual subjects. The impact of the global SNR on measures of individual reliability is assessed, as well as the impact on the $ICC_{between}$.

Materials and methods

Subjects

12 Healthy subjects (6 males, 6 females; mean ± SD age, 22.1 ± 1.75 years) recruited from the University of Utrecht participated in the experiment. None of the subjects had any signs of present or past major psychiatric illnesses according to the Mini-International Neuropsychiatric Interview (Sheehan et al., 1998). The subjects had no previous experience with fMRI or with the oculomotor task. A history of major neurological illness resulted in exclusion from the experiment, as did metal implants. All subjects gave informed consent for participation (approved by the Human Ethics Committee of the University Medical Center Utrecht). All were right handed according to the Edinburgh Handedness inventory (Oldfield, 1971). The test and retest took place at the same time in the evening, with an interval of 1 week. All subjects were non-smokers and were asked to make sure that coffee intake and amount of sleep was equal for both days of scanning.

Scanning protocol

All images were obtained with a Philips ACS-NT 1.5 T MRI scanner (Philips Medical Systems, Best, The Netherlands) with fast gradients (PT6000). The head was held in place with a strap and with padding. Functional images were acquired in transverse orientation, and encompassed the whole brain except for the cerebellum, the orbitofrontal cortex, and the inferior temporal cortex. For functional scans, a navigated 3D-PRESTO pulse sequence (Ramsey et al., 1998; van Gelderen et al., 1995) was used with following parameters: TE 37 ms, TR 24 ms, flip angle 9.5°, matrix 48 × 64 × 24, FOV 192 × 256 × 96 mm, voxel size 4 mm isotropic, scan duration 1.49 s per 24-slice volume. Immediately after functional scans, an additional PRESTO scan of the same volume of brain tissue was acquired with a high flip-angle (30°, FA30) for the image coregistration routine (see below). A T1-weighted structural image of the whole brain was acquired at the end of both sessions.

Task design

The fMRI design used a PC, a rear projection screen and a video-projector system for presentation. All stimuli were projected in white on a dark background. All events were time locked to the fMRI scans. Instructions were given verbally, prior to the start of the experiment. With the aid of a laptop, a limited number of test trials were presented before scanning, until subjects indicated they understood the task. The design consisted of two tasks, i.e., prosaccades and antisaccades, which had identical stimuli. Whether subjects were requested to make prosaccades or antisaccades depended on a short summary of the instructions at the beginning of each new block of ten stimuli. Instructions were presented for a duration of three scans, followed by a 6 scan period of central fixation. Each new trial started with the disappearance of a fixation

cross (0.9° visual angle) at central view. After a 0.2-s gap period, a square (0.9° visual angle) was presented semi-random 8.7° to the left or right of central fixation. If the instructions were to make prosaccades, subjects had to make a saccade towards the square as quickly as possible. If the instructions were to make antisaccades, subjects had to avoid an automatic eye movement towards the square, and instead make a saccade towards the opposite direction. The square was extinguished after 3.24 s, simultaneously with the reappearance of the fixation cross at central view. This signaled the subjects to refixate in the center of the screen. A new stimulus was triggered 10.16 s after central refixation, thereby generating a fixed stimulus interval of 13.4 s giving stimulus-related BOLD signal time to return to baseline (Bandettini and Cox, 2000). Stimulus-related changes in BOLD signal were thus measured relative to fixating in the center or in the periphery. The long intertrial interval with specific event delays was used to avoid high correlations between the prosaccade or antisaccade and the saccade that returned the gaze back to the center (Raemaekers et al., 2005, 2006b). There were four blocks per task making a total of eight, which were orderly alternated. Each subject made 40 prosaccades and 40 antisaccades, and 792 functional scans were acquired in a single session of 19 min and 39 s (Fig. 1).

Analysis

All preprocessing steps were done using SPM2 (http://www.fil.ion.ucl.ac.uk/spm/). After realignment, the functional scans were coregistered to the FA30 volume, using the first functional volume as a source. The structural scan was also coregistered to the FA30-scan, thereby providing spatial alignment between the structural scan and the functional volumes. Normalization parameters were estimated using the MNI T1-standard brain as template (Collins et al., 1994), and the coregistered T1 volume as a source. All functional scans were then normalized and resliced to a $4 \times 4 \times 4$-mm resolution. A 3D Gaussian filter (8-mm full width at half max) was applied to all fMRI volumes.

The data of each scanning session were submitted to two separate multiple regression analyses using IDL (Research Systems Inc., Boulder, USA). The first design matrix contained five factors that represented prosaccades, antisaccades, refixation in the center during prosaccades, refixation in the center during antisaccades, and reading of the instructions respectively. The factors were based on short (1-ms event) box cars that were temporally aligned with the stimulus events. The events were convolved with a predefined hemodynamic response function (canonical HRF of SPM2), and interpolated to a scan duration scale (Friston et al., 1995). Correct and incorrect trials were not separately modeled in the design matrix.

A second multiple regression analysis was done to estimate the SNR of the BOLD response irrespective of shape, and also to calculate average time series for prosaccade and antisaccade trials. The second design matrix contained 8 finite impulse response (FIR) functions per condition (leaving the first scan of each trial as reference). An estimate of the variance induced by the BOLD response was made for each voxel by calculating the F values for the F-contrasts including the FIR set of prosaccades, and for the contrast including the FIR set of antisaccades. Both design matrices contained factors for low frequency noise, i.e., the mean signal intensity of each scan, and 88 discrete cosine functions forming a high pass filter with a cut-off at 3.73×10^{-2} Hz to correct for low-frequency scanner and physiological artifacts, differences in baseline activation between conditions, and low-frequency signal changes as a result of head motion.

The t-statistics of the relevant contrasts (e.g., prosaccades, antisaccades, and antisaccades vs. prosaccades) were calculated for every voxel. Subsequently, average t-values for each contrast were calculated in eight predefined ROIs including V1, V2/V5, parietal cortex, the frontal eye fields (FEF), the insula, the striatum, the supplementary eye fields (SEF), and the dorsolateral prefrontal cortex (DLPFC) (Fig. 2). There was no DLPFC ROI for the contrast between antisaccades and prosaccades. The ROIs were based on the three statistical group maps of 36 subjects that did the same task in previous experiments (Raemaekers et al., 2005, 2006b). Systematic differences in brain activation between the sessions were estimated for the three contrasts separately by comparing average t-values in the ROIs with a multivariate (8 ROIs for antisaccades and prosaccades, and 7 ROIs for the contrast between antisaccades and prosaccades) repeated measures (session 1 and session 2) MANOVA. The voxel wise results for the group were analyzed for each session separately using one sample t-tests on the t-volumes for prosaccades, antisaccades, and the contrast between antisaccades and prosaccades.

Analyses of reliability

For all measures of reliability and sensitivity of individual results, a mask was used that contained the areas of the brain that were scanned in all subjects, and excluded voxels that were located in the ventricles based on the individual normalized T1 scan.

Overlap in activation

The measure of overlap in activation was used to assess the test–retest reliability of the brain activation of individual subjects, as well as the test–retest reliability of the brain activation found for the group-wise comparisons. This was done by calculating the relative amount of volume overlap in activation between the two sessions for the t-volumes of individual subjects, as well as for the t-volumes of the group-wise comparisons. Overlap was estimated for prosaccades, antisaccades, and the contrast between antisaccades and prosaccades separately. The overlap between the sessions (R^{12}_{overlap})

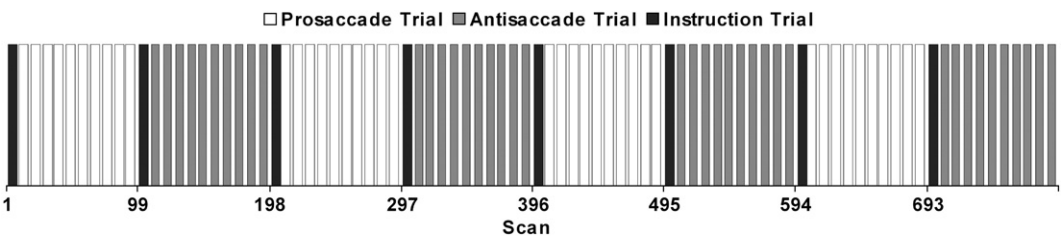

Fig. 1. Schematic representation of the stimulus presentation. Bars indicate individual trials. The time axis is in scan durations (1.49 s).

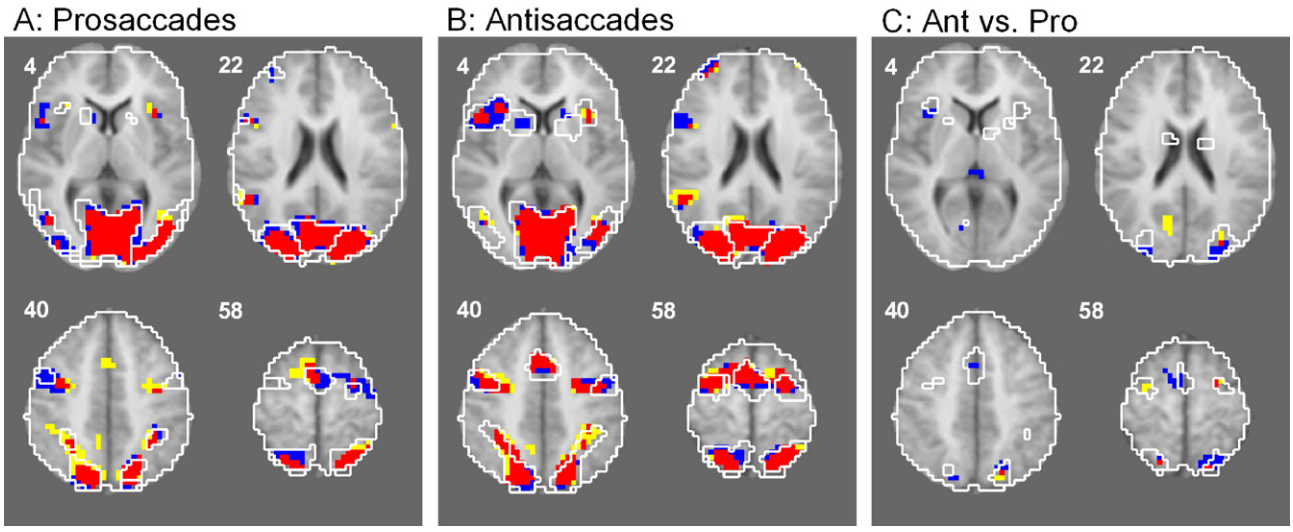

Fig. 2. Voxels that were active in the group comparisons during either session 1 (blue), session 2 (yellow), or both sessions (red), superimposed on the average anatomical scan. The numbers displayed on the top left corner of each slice correspond to the Talairach z coordinates. Only the most informative slices are shown. Colored voxels represent significant effects at $p<0.001$ (uncorrected). Panel A depicts the results for prosaccades, panel B for antisaccades, and panel C for antisaccades vs. prosaccades. The white lines encircle the ROIs that are defined based on the results of previous studies and involve V1, V2/V5, parietal, FEF, the insula, striatum, the SEF, and the DLPFC.

was calculated by using the formula proposed by Rombouts et al. (1998) and Machielsen et al. (2000):

$$R^{12}_{\text{overlap}} = \frac{2*V_{\text{overlap}}}{V_1 + V_2}$$

V_1 and V_2 denote the number of voxels in the whole t-volume passing the statistical threshold in session 1 and session 2 respectively, and V_{overlap} the number of voxels that pass the threshold in both whole t-volumes. The R^{12}_{overlap} can range from 0 (no overlap) to 1 (perfect overlap), and is a descriptive statistic for the ratio of the number of voxels that are active in both sessions and the total number of active voxels. For estimating the R^{12}_{overlap}, a statistical threshold of $p<0.001$ (uncorrected) was used.

Intraclass correlation within measurements (ICC_{within})

Another measure for estimation of reliability of fMRI results is the intraclass correlation of contrast t-values for pairs of activation maps. Like the R^{12}_{overlap}, this measure can be used to assess the test–retest reliability of both individual subject data and the group-wise results. For this purpose, a two-way random ICC for absolute agreement between the measurements was used (R^{12}_{overlap} also estimates absolute agreement) (Shrout and Fleiss, 1979). In contrast to the R^{12}_{overlap}, the $\text{ICC}_{\text{within}}$ is based on all the voxels in the brain. Therefore, it is not dependent on the choice for a particular significance threshold. On the negative side, inclusion of all voxels means that the value is also determined by brain areas that are not involved in the task. All t-values in the t-volume of session 1 are correlated with the t-values in the t-volume of session 2 ($\text{ICC}_{\text{within}}$). This was done for all contrasts (prosaccades, antisaccades, and antisaccades vs. prosaccades) and for the group t-volumes, as well as for the individual t-volumes.

$$\text{ICC}_{\text{within}} = \frac{\text{MS}_{\text{between}} - \text{MS}_{\text{error}}}{\text{MS}_{\text{between}} + \text{MS}_{\text{error}} + 2*(\text{MS}_{\text{column}} - \text{MS}_{\text{error}})}$$

$\text{MS}_{\text{between}}$ is the mean square of the variance in t-values between voxels. $\text{MS}_{\text{column}}$ is the mean square of the systematic (column) differences in voxel t-values between the two sessions. MS_{error} is equal to the mean square of the within voxel variance (over sessions) after removal of the systematic session (column) variance. For averaging the intraclass correlation coefficients across subjects for two groups, Fisher's z-transformation can be used on the individually estimated $\text{ICC}_{\text{within}}$ before group-wise comparisons:

$$z' = \left(\frac{1}{2}\right)\log\left(\frac{1+R^{12}}{1-R^{12}}\right)$$

Intraclass correlation between measurements ($ICC_{between}$)

The measures of R^{12}_{overlap} and $\text{ICC}_{\text{within}}$ can be used to assess reliability within a single measurement (either an individual or a group-wise measurement). Accurate classification or phenotyping of subjects does not only depend on the reliability of individual measurements, but also on the variance between subjects. Both these facets are incorporated in the intraclass correlation between measurements ($\text{ICC}_{\text{between}}$). The two observations in this experiment differ in a systematic way, due to effects of previous exposure to the experiment in the second session. As endophenotypes are normally measured based on a single session, systematic differences between the measurements are deemed irrelevant in this experiment. Therefore, the two-way random ICC for consistency was chosen (Shrout and Fleiss, 1979). The $\text{ICC}_{\text{between}}$ is calculated for two within-subject measurements as:

$$\text{ICC}_{\text{between}} = \frac{\text{MS}_{\text{between}} - \text{MS}_{\text{error}}}{\text{MS}_{\text{between}} + \text{MS}_{\text{error}}}$$

$\text{MS}_{\text{between}}$ and MS_{error} are the mean square for between-subject and error variance, respectively. The MS_{error} is equal to the mean square of the within-subject variance after removal of the systematic session (column) variance. The $\text{ICC}_{\text{between}}$ thus represents the ratio

of between-subject variance to total variance (without systematic session variance). The $ICC_{between}$ was determined for the average *t*-values in all the predefined ROIs (Fig. 2).

Calculation of the sensitivity ($M_{|t|}$)

Although it is common in fMRI to calculate the SNR on a voxel by voxel basis, the SNR within a voxel is also under the influence of global factors that affect the entire brain. A global difference in the SNR between the subjects causes differences in statistical sensitivity between subjects. When the statistical sensitivity of an individual measurement is low, e.g., due to motion artifacts, scanner noise or physiological factors, the *t*-values in the statistical maps of individual measurements will be low and therefore irreproducible. This will result in a low $R^{12}_{overlap}$, and a low ICC_{within}. To estimate the average amplitude of the *t*-values of the individual measurements, for each subject and session, the average absolute *t*-value during prosaccades and antisaccades was calculated.

$$M_{|t|} = \frac{1}{2*n} * \sum_{i=1}^{n} \left(|t_{i;\text{prosaccades}}| + |t_{i;\text{antisaccades}}| \right)$$

In the formula *t* depicts the *t*-value per voxel in the prosaccades and antisaccades contrast, and *n* the number of voxels. $M_{|t|}$ represents the total amount of stimulus related changes in the brain activation relative to the noise, for prosaccades and antisaccades combined. The $ICC_{between}$ of $M_{|t|}$ was calculated to estimate the stability of statistical sensitivity of measurements in individuals. Subsequently, the $M_{|t|}$ for each subject was averaged over the sessions and was correlated with $R^{12}_{overlap}$ and ICC_{within} to estimate to what extent individual reliability related to statistical sensitivity. Furthermore, it was estimated whether intersubject differences in $M_{|t|}$ could underlie part of the between-subject variance when estimating the $ICC_{between}$ for different ROIs and task contrasts. Therefore, the $ICC_{between}$ was recalculated after removal of between-subject variance that could be explained by the average $M_{|t|}$ over the two sessions. A secondary measure of sensitivity was included to assess the individual statistical sensitivity irrespective of the shape of the BOLD response.

$$M_{\sqrt{F}} = \frac{1}{2*n} * \sum_{i=1}^{n} \left(\sqrt{F_{i;\text{prosaccades}}} + \sqrt{F_{i;\text{antisaccades}}} \right)$$

In this formula, $\sqrt{F}$ depicts the square root of the *F*-value per voxel corresponding to the two FIR sets, and *n* the number of voxels. $M_{\sqrt{F}}$ represents the total amount of stimulus related changes in the brain, irrespective of a model of the BOLD response. To assess the extent to which intersubject differences in $M_{|t|}$ could arise as a result of differences in the amplitude of the BOLD response relative to the noise, instead of systematic deviations from the shape of the model, the $M_{\sqrt{F}}$ and $M_{|t|}$ measure were correlated.

Eye movements

Eye movements were recorded during the entire oculomotor-task using an MR-compatible eye tracker (Cambridge Research Systems Ltd., Rochester, UK (Kimmig et al., 1999)) in combination with Labview (National Instruments Corporation, Austin, USA) acquisition software on a PC with a multifunctional I/O Board (National Instruments Corporation, Austin, USA). This acquisition-PC was linked to the stimulus-PC by a parallel cable to synchronize the eye recordings and the task presentation. Calibration and adjustment of the sensor were done during a 5-min period prior to scanning. The sample frequency of the recording was 500 Hz. For each saccade in the time window of 200 ms before until 600 ms after the stimulus presentation, the latency and the direction were determined using a custom non automated analysis program in IDL (Research Systems Inc., Boulder, USA). Reliability was estimated by calculating the $ICC_{between}$ of the different behavioral measures.

Results

Systematic changes

There was a significant multivariate effect of retesting on activation levels in the predefined ROIs (V1, V2/V5, parietal cortex, the FEF, the insula, the striatum, the SEF, and the DLPFC) for the prosaccade condition ($F_{(8,4)}=11.43$; $p=0.016$) (Fig. 3). This effect was constituted by signal reductions throughout the brain, mostly in V1 ($F_{(1,11)}=6.40$; $p=0.028$) and the SEF ($F_{(1,11)}=6.02$; $p=0.032$). There was no significant multivariate effect of retesting on the antisaccade condition ($F_{(8,4)}=0.65$; $p=0.72$), nor for the contrast between antisaccades and prosaccades ($F_{(7,5)}=0.32$; $p=0.91$), The average time courses for prosaccade and antisaccade trials over the two sessions in the eight ROIs can been seen in Fig. 4.

Overlap ratios ($R^{12}_{overlap}$)

The $R^{12}_{overlap}$ for the group *t*-volumes for prosaccades, antisaccades, and antisaccades vs. prosaccades were 0.76, 0.81, and

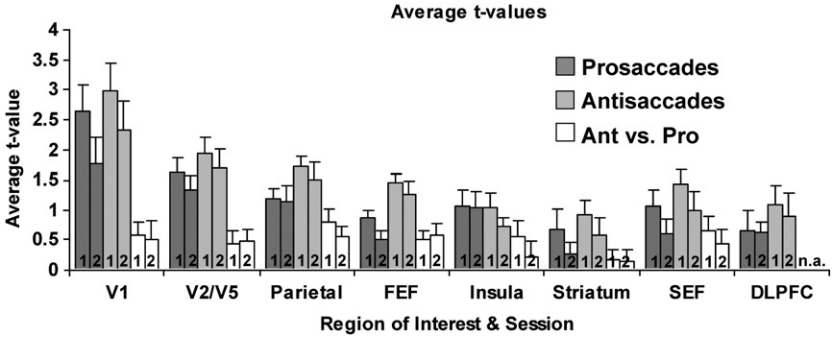

Fig. 3. Average *t*-values for the three contrasts in the relevant ROIs (Fig. 2). The number at the bottom of each bar signifies the session number. Bars indicate standard errors. There was no DLPFC ROI for the contrast between antisaccades and prosaccades.

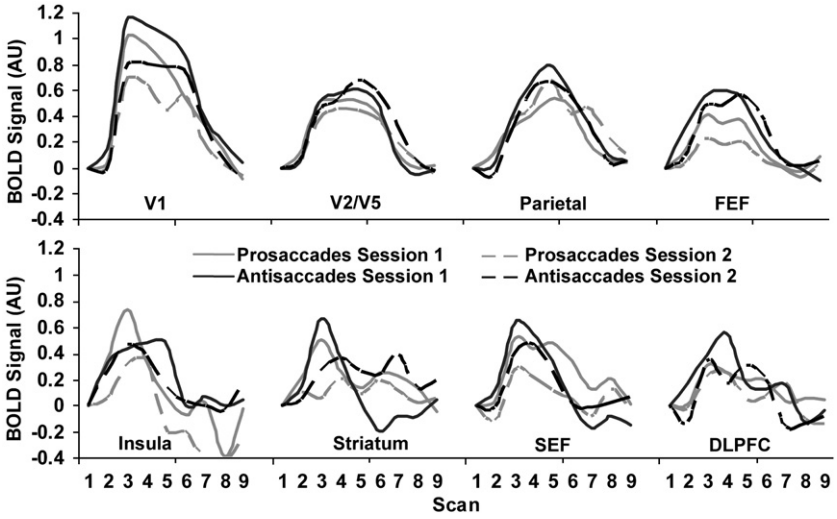

Fig. 4. Averaged BOLD responses related to prosaccade and antisaccade trials for session 1 and session 2 in the eight predefined ROIs (Fig. 2). Because of the normalization of values in multiple regression analysis, the BOLD response amplitude is presented in arbitrary units (AU).

0.27 respectively (Fig. 2). Thus, with the current number of subjects, group results for prosaccades and antisaccades are highly reproducible. The lower overlap for antisaccades vs. prosaccades activation is probably the result of relatively less statistical power of individual measurements. The averages of the $R^{12}_{overlap}$ of the individual subjects were 0.32 (±0.23) for prosaccades, 0.41 (±0.28), for antisaccades, and 0.08 (±0.08) for the contrast between antisaccades and prosaccades. The overlap ratios could substantially differ between subjects (00–0.72 for prosaccades, 0.05–0.84 for antisaccades, 0.00–0.24 for antisaccades vs. prosaccades).

Intraclass correlation within measurements (ICC_{within})

The *t*-values of the group results of session 1 are plotted against the *t*-values of the group results of session 2 in Fig. 5 for the three contrasts. The intraclass correlations of *t*-values between the sessions were all significant ($ICC_{within}=0.88$; $p<0.001$ for prosaccades, $ICC_{within}=0.88$; $p<0.001$ for antisaccades, and $ICC_{within}=0.43$; $p<0.001$ for antisaccades vs. prosaccades) (Fig. 5). The average transformed correlations (*z* values) were 0.52 (±0.30) for prosaccades, 0.69 (±0.30) for antisaccades, and 0.16 (±0.15) for antisaccades vs. prosaccades. For individual subjects, the ICC_{within} differed substantially (0.02–0.77 for prosaccades, 0.29–0.79 for antisaccades, −0.08 to 0.40 for antisaccades vs. prosaccades) (see Fig. 6 for two examples).

Intraclass correlations between measurements ($ICC_{between}$)

Result for the ROI-based $ICC_{between}$ estimation are similar (Table 1A). Areas of reasonable and good reliability can be found mostly in occipital areas and the SEF during prosaccades and antisaccades. Results in other areas are mixed, and differ between prosaccades and antisaccades. Reliability for brain activation for the contrast between antisaccades and prosaccades is poor in general.

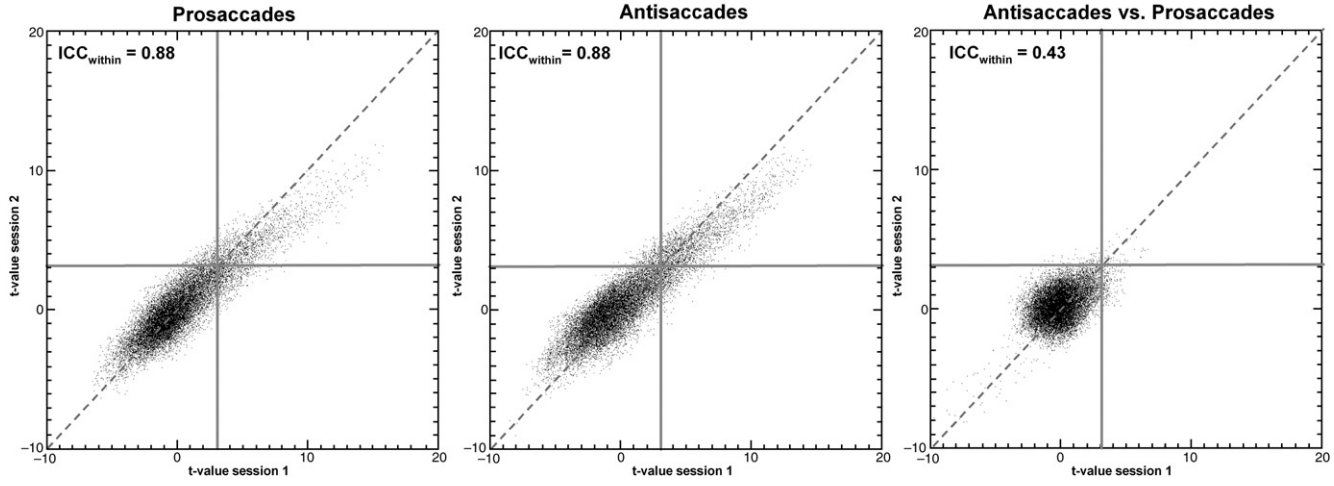

Fig. 5. For the three contrasts, the *t*-values of the group results of the first session plotted against the *t*-values for the second session. Grey lines indicate the statistical threshold that was used to calculate the $R^{12}_{overlap}$.

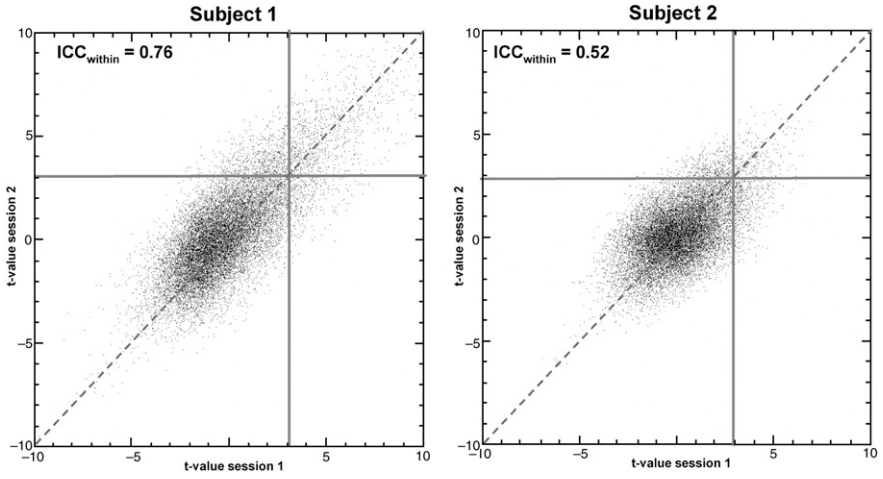

Fig. 6. For two individuals, the *t*-values for antisaccades in the first session plotted against the *t*-values for antisaccades in the second session. Grey lines indicate the statistical threshold that was used to calculate the $R^{12}_{overlap}$.

Statistical sensitivity

The reliability of the $M_{|t|}$ was high over the two sessions ($ICC_{between}=0.80$; $p<0.001$, 95% confidence interval=0.45–0.94) (Fig. 7). When averaged over the sessions, there were large intersubject differences in $M_{|t|}$ (ranging between 0.98 and 1.69). Further analysis of the $M_{|t|}$ value revealed that its magnitude was largely independent of the specific contrast (prosaccades or antisaccades), as the correlation between the $M_{|t|}$ based solely on prosaccades and the $M_{|t|}$ based solely on antisaccades was high ($r=0.87$; $p<0.001$). To assess up to what extent individual differences in sensitivity could be explained by systematic deviations of the empirical HRF from the shape of the model, we correlated the $M_{|t|}$ value with the $M_{\sqrt{F}}$ measure (which has no assumption on the shape of the HRF due to the use of FIRS). The distribution of $M_{\sqrt{F}}$ was very similar to the $M_{|t|}$ distribution (Pearson $r=0.95$: $p<0.001$; Spearman's $\rho=0.94$; $p<0.001$). This suggests that intersubject differences in $M_{|t|}$ largely arise from intersubject differences in size of the BOLD response (relative to the noise), instead of systematic intersubject differences in deviation from the shape of the model.

The correlations between the average $M_{|t|}$ and the $R^{12}_{overlap}$, were significant for prosaccades and antisaccades ($r=0.85$; $p<0.001$ for prosaccades, $r=0.94$; $p<0.001$ for antisaccades), but not for the contrast between antisaccades and prosaccades ($r=0.27$; $p=0.40$). The correlations between the $M_{|t|}$ and z were significant for all contrasts ($r=0.91$; $p<0.001$ for prosaccades, $r=0.90$; $p<0.001$ for antisaccades, $r=0.73$; $p=0.007$ for antisaccades vs. prosaccades). These correlations show that differences between subjects in overlapping activation over the two sessions ($R^{12}_{overlap}$), and differences in correlations of the individual *t*-values between subjects (z) can be largely explained by intersubject differences in $M_{|t|}$. High correlations between $M_{|t|}$ and average *t*-values in the ROIs show that a large proportion of the between-subject variance

Table 1
$ICC_{between}$ estimates with *P* values for average *t*-values in the seven predefined ROIs (Fig. 2) for prosaccades, antisaccades, and the contrast between antisaccades and prosaccades

	n	Prosaccades		Antisaccades		Antisaccades vs. Prosaccades	
		$ICC_{between}$	*P*	$ICC_{between}$	*P*	$ICC_{between}$	*P*
(A)							
V1	12	0.71*	0.00	0.85*	0.00	0.30	0.16
V2/V5	12	0.76*	0.00	0.80*	0.00	0.46	0.05
Parietal	12	0.68*	0.00	0.28	0.16	0.29	0.16
FEF	12	0.31	0.14	0.49*	0.04	0.39	0.09
Insula	12	−0.06	0.57	0.35	0.11	−0.25	0.80
Striatum	12	0.45	0.05	0.19	0.26	−0.33	0.86
SEF	12	0.75*	0.00	0.55*	0.02	0.11	0.35
DLPFC	12	0.54*	0.03	0.16	0.29	n.a.	n.a.
(B)							
V1	12	0.05	0.43	0.58*	0.02	0.29	0.16
V2/V5	12	0.17	0.29	0.45	0.06	0.40	0.08
Parietal	12	0.25	0.19	0.03	0.45	0.25	0.19
FEF	12	0.04	0.45	0.27	0.18	0.38	0.09
Insula	12	−0.06	0.58	0.33	0.13	−0.26	0.80
Striatum	12	0.16	0.29	0.08	0.39	−0.34	0.88
SEF	12	0.52*	0.03	0.20	0.25	0.02	0.46
DLPFC	12	0.54*	0.03	0.16	0.29	n.a.	n.a.

A: $ICC_{between}$ estimates based on uncorrected average *t*-values; B: $ICC_{between}$ estimates after removal of the between subject variance that could be explained by $M_{|t|}$.

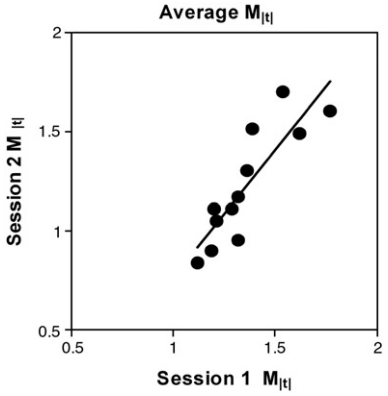

Fig. 7. Scatter plot of the $M_{|t|}$ during session 1, against the $MS_{|t|}$ during session 2.

in the regional brain activation could be explained by $M_{|t|}$ (Table 2). After removing the between-subject variance from the ROI activation that could be explained by $M_{|t|}$ using a linear regression, the $ICC_{between}$ scores for each region and contrast were reassessed. This correction resulted in a large reduction in the $ICC_{between}$ in many ROIs throughout the brain (Table 1B).

Results of behavioral measures

When comparing the behavioral measures of the two sessions, there were no significant differences in performance (Table 3). The stability of the onset latencies of both correct prosaccades and correct antisaccades was high (prosaccades (Fig. 8A); $ICC_{between}$ = 0.86, 95% confidence interval = 0.56–0.96, $p < 0.001$; antisaccades (Fig. 8B); $ICC_{between}$ = 0.91, 95% confidence interval = 0.67–0.98, $p < 0.001$). The error rates for antisaccades were less stable over the two sessions (Fig. 8C) ($ICC_{between}$ = 0.47, p = 0.07, 95% confidence interval = −0.18–0.84). The behavioral data in two subjects in session 1 were partially unusable due to the loss of the eye position signal as a result of subject movement. These two subjects had an error rate for antisaccades 70% and 61% in session 2. As these are much larger than the average antisaccade error rate, the loss of the data may have resulted in an underestimation of the between-subject variance, and subsequently the $ICC_{between}$. Errors during prosaccades were too sporadic to make a reliable $ICC_{between}$ estimate.

Discussion

In this experiment, the test–retest reliability of individual and group-wise fMRI activation maps during prosaccades, antisaccades, and the contrast between antisaccades and prosaccades was assessed. The group activation maps for prosaccades and antisaccades were very similar across the two sessions. The reliability of the group map for the contrast between antisaccades and prosaccades was considerably lower. Reliability of individual measurements was variable between subjects, and was highly correlated with the measure of statistical sensitivity, and thus with intersubject differences in the global temporal SNR. In addition, high correlations between our sensitivity measure and t-values in many different ROIs, indicate that intersubject differences in regional brain activation could be explained to a large extent by between-subject differences in the global SNR. This indicates that the strongest determinant of differences between subjects was the overall SNR, as opposed to regionally selective strength of activation. In other words, some subjects display large regions of activations whereas others display small regions, and this feature is reproducible across sessions. Behavioral measures had moderate to good reliability.

As reported previously, fMRI group-wise results can be highly reproducible (Casey et al., 1998). Reliability of the group results for the contrast between antisaccades and prosaccades could still be considerably improved, however. In comparison to prosaccades and antisaccades, the contrast between antisaccades and prosaccades had lower effect sizes of individual measurements, which probably caused the lesser correlation and overlap of activation. Higher field strengths may be required to detect more robust activation in both individual and group-wise results for the contrast between antisaccades and prosaccades, as is observed in other studies (Brown et al., 2006; Connolly et al., 2002; Curtis and D'Esposito, 2003; Desouza et al., 2003; Ford et al., 2005).

When comparing the group-wise activation between the two sessions, there were decreases in activation from session 1 to session 2 for prosaccades. Reductions in activation have been associated with learning. It is well known that practice can lead to a reduction in functional activation over time (Chein and Schneider, 2005). Practice induced reductions in activation are thought to reflect functional trimming of neuronal ensembles that are involved in the task (Ramsey et al., 2004). However, the global changes in the fMRI signal that we found were not accompanied by changes in task performance. The absence of substantial learning effects may indicate that the global signal change between the sessions is related to familiarity with the fMRI procedure, in that subjects are less aroused in the second session.

Before estimating the $ICC_{between}$, the reliability of individual measurements was assessed, as this is one of the cornerstones of the reproducibility of between-subject variation. We found a large variation in individual reliability, concerning both the overlap ratios ($R^{12}_{overlap}$) and the intraclass correlation of t-values within subjects (ICC_{within}). This variation in reliability could be largely explained by differences in SNR in individual measurements. Good individual reliability could thus be obtained if the SNR of the measurement was high enough. A better SNR increases statistical sensitivity (i.e., high t-values in individual t-maps), indicating that individual reliability is mostly a matter of statistical sensitivity and not variability in brain activation between sessions. Hence, although fMRI reliability will be different for different scanning techniques (e.g., EPI or PRESTO), 1.5 or 3 T, voxel sizes, specific task and task design, and even preprocessing and analysis techniques (Smith et al., 2005), it is the difference in statistical sensitivity that is probably most important. The average overlap ratios that we estimated for the three contrasts were only moderate compared to the overlap ratios that

Table 2
Correlation estimates with P values for the correlation between $M_{|t|}$ and the average t-values in the predefined ROIs (Fig. 2) for prosaccades, antisaccades, and antisaccades vs. prosaccades

	n	Prosaccades		Antisaccades		Antisaccades vs. Prosaccades	
		R	P	R	P	R	P
V1	12	0.90*	0.00	0.84*	0.00	0.08	0.81
V2/V5	12	0.90*	0.00	0.84*	0.00	0.37	0.23
Parietal	12	0.83*	0.00	0.64*	0.03	−0.26	0.41
FEF	12	0.66*	0.02	0.64*	0.02	0.14	0.67
Insula	12	−0.11	0.74	−0.19	0.56	−0.13	0.68
Striatum	12	0.69*	0.01	0.44	0.15	−0.18	0.57
SEF	12	0.74*	0.01	0.75*	0.00	0.40	0.20
DLPFC	12	0.63*	0.03	0.41	0.19	n.a.	n.a.

Table 3
Summary of the behavioral results of prosaccades and antisaccades for the two sessions

	n	Session 1		Session 2		Paired t
		M	(SD)	M	(SD)	
Prosaccade latencies	11	177.42	19.00	181.91	22.34	−1.35
Prosaccade %errors	11	0.70	2.32	0.79	1.78	−0.15
Antisaccade latencies	10	243.24	37.97	238.81	35.39	0.88
Antisaccade %errors	10	16.36	9.27	18.42	15.14	−0.51

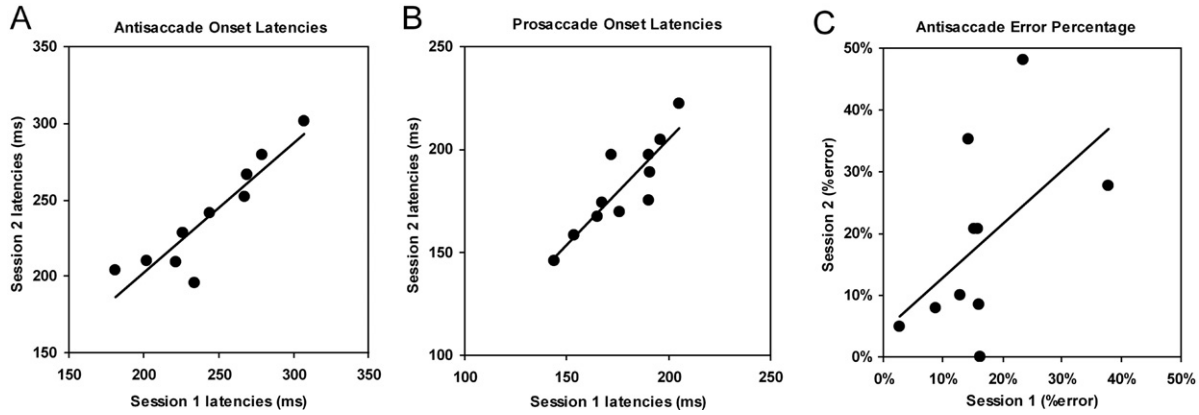

Fig. 8. Scatter plots of the behavioral measures of the first session, against the behavioral measures of the second session. (A) Saccade onset latencies for correct prosaccades; (B) saccade onset latencies for correct antisaccades; (C) error percentage during antisaccades.

have been reported in the literature for other fMRI paradigms (Machielsen et al., 2000; Rombouts et al., 1998; Specht et al., 2003). This is probably due to the fact that the other tasks were primarily visual, and were presented in a blocked design, a design type that is statistically very powerful. The current design is sparse event related, which may result in lower statistical power, and thus a lower overlap in activation.

A large portion of the between-subject variance in brain activation in the predefined ROIs can be explained by differences in the global SNR of the individual measurements. Although the current sample size is too small to provide precise estimates, the $ICC_{between}$ scores of this corrected ROI-based activation are moderate at best. We found evidence that the individual differences in the global SNR are quite stable over time ($ICC_{between}=0.80$). In addition, the SNR was also stable over the prosaccade and antisaccade conditions ($r=0.87$). Intersubject differences in SNR were also detected irrespective of the shape of the hemodynamic response, suggesting that they are primarily a result of differences in amplitude of the hemodynamic response (relative to the noise) and not the shape. Intersubject differences in the global SNR are not readably interpretable. They could be linked to intersubject differences in subjective attentional demand of the task (Adler et al., 2001). On the other hand, they could also be related to global differences, e.g., physiological noise, size of the hemodynamic response, motion artifacts, etc. Future studies should address the origin of individual differences in the SNR in fMRI measurements.

As a measure of the individual statistical sensitivity of the measurement, we used the mean of all absolute t-values ($M_{|t|}$) during prosaccades and antisaccades. One could argue that regionally specific differences in brain activation will also result in differences in the $M_{|t|}$. By correcting the average t-values in the ROIs before $ICC_{between}$ estimation, this regional between-subject variation would be removed. However, the between-subject differences in $M_{|t|}$ that we found could be so large (ranging between 0.98 and 1.69), that they can hardly be explained by regionally specific differences in activation. In addition, as $M_{|t|}$ correlates highly with the average t-values of nearly all areas of the visual and oculomotor system during prosaccades and antisaccades, it is clear that much of the individual variation in the ROIs has a common source, and that $M_{|t|}$ reflects this common source. The contribution of regionally specific differences in brain activation to $M_{|t|}$ are thus at best minimal. However, to completely remove the contribution of regionally specific variations, the measure of statistical sensitivity should be generated independently of the measurement. Recently proposed methods for individual calibration of the BOLD response using a breath holding challenge may become very useful in this context (Thomason et al., 2007).

There was no performance increase on the task between session 1 and session 2. A previous study reported learning effects on the number of errors during the antisaccade task (Ettinger et al., 2003), but these could have been largely cancelled in this study due to intermingling of prosaccade and antisaccade blocks (Dyckman and McDowell, 2005). As for the reproducibility of behavioral results, we found high test–retest reliability of onset latencies of both prosaccades and antisaccades, similar to what is reported in the literature (Ettinger et al., 2003; Harris et al., 2006; Klein and Berg, 2001; Roy-Byrne et al., 1995). Previous reports on the reliability of the percentage of erroneous distractions during antisaccades are less consistent (ranging between −0.30 and 0.89). This measure may be more dependent on task specific variations (e.g., gap/overlap) between these studies which may affect intersubject variability. More difficult paradigms tend to increase the between-subject variability and thereby the $ICC_{between}$. The $ICC_{between}$ score of 0.47 that was found in this experiment is relatively small and just not significant. Due to the loss of data of two subjects that had a very high error rate, the between-subject variation may have been underestimated, which may have resulted in a lower $ICC_{between}$. A larger sample size is needed to make a more precise estimate of the reliability of distractibility in this particular design.

In summary, group-wise t-maps of prosaccades and antisaccades can be highly reliable in a group with twelve subjects, but for activation maps with smaller effect sizes, like for the contrast between antisaccades and prosaccades, probably more subjects are required. As for individual maps, more statistical power is needed to make a more reliable estimate of the amplitude of the BOLD response, especially in individuals that have a poor SNR. Large differences exist between subjects in the global SNR, and the global SNR is a good predictor for reproducibility of individual data sets. Future studies on test–retest reliability should address differences in the global SNR between subjects, before estimating the ROI based $ICC_{between}$. Endophenotyping based on brain function could benefit from incorporating intersubject differences in SNR as well.

References

Adler, C.M., Sax, K.W., Holland, S.K., Schmithorst, V., Rosenberg, L., Strakowski, S.M., 2001. Changes in neuronal activation with increasing attention demand in healthy volunteers: an fMRI study. Synapse 42, 266–272.

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, 1000–1006.

Bandettini, P.A., Cox, R.W., 2000. Event-related fMRI contrast when using constant interstimulus interval: theory and experiment. Magn. Reson. Med. 43, 540–548.

Brown, M.R., Goltz, H.C., Vilis, T., Ford, K.A., Everling, S., 2006. Inhibition and generation of saccades: rapid event-related fMRI of prosaccades, antisaccades, and nogo trials. NeuroImage 33, 644–659.

Calkins, M.E., Iacono, W.G., Curtis, C.E., 2003. Smooth pursuit and antisaccade performance evidence trait stability in schizophrenia patients and their relatives. Int. J. Psychophysiol. 49, 139–146.

Calkins, M.E., Curtis, C.E., Iacono, W.G., Grove, W.M., 2004. Antisaccade performance is impaired in medically and psychiatrically healthy biological relatives of schizophrenia patients. Schizophr. Res. 71, 167–178.

Callicott, J.H., Egan, M.F., Mattay, V.S., Bertolino, A., Bone, A.D., Verchinksi, B., Weinberger, D.R., 2003. Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. Am. J. Psychiatry 160, 709–719.

Casey, B.J., Cohen, J.D., O'Craven, K., Davidson, R.J., Irwin, W., Nelson, C.A., Noll, D.C., Hu, X., Lowe, M.J., Rosen, B.R., Truwitt, C.L., Turski, P.A., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8, 249–261.

Chein, J.M., Schneider, W., 2005. Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. Brain Res. Cogn. Brain Res. 25, 607–623.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 18, 192–205.

Connolly, J.D., Goodale, M.A., Menon, R.S., Munoz, D.P., 2002. Human fMRI evidence for the neural correlates of preparatory set. Nat. Neurosci. 5, 1345–1352.

Curtis, C.E., D'Esposito, M., 2003. Success and failure suppressing reflexive behavior. J. Cogn. Neurosci. 15, 409–418.

Desouza, J.F., Menon, R.S., Everling, S., 2003. Preparatory set associated with pro-saccades and anti-saccades in humans investigated with event-related fMRI. J. Neurophysiol. 89, 1016–1023.

Dyckman, K.A., McDowell, J.E., 2005. Behavioral plasticity of antisaccade performance following daily practice. Exp. Brain Res. 162, 63–69.

Ettinger, U., Kumari, V., Crawford, T.J., Davis, R.E., Sharma, T., Corr, P.J., 2003. Reliability of smooth pursuit, fixation, and saccadic eye movements. Psychophysiology 40, 620–628.

Ettinger, U., Picchioni, M., Hall, M.H., Schulze, K., Toulopoulou, T., Landau, S., Crawford, T.J., Murray, R.M., 2006. Antisaccade performance in monozygotic twins discordant for schizophrenia: the Maudsley twin study. Am. J. Psychiatry 163, 543–545.

Ford, K.A., Goltz, H.C., Brown, M.R., Everling, S., 2005. Neural processes associated with antisaccade task performance investigated with event-related fMRI. J. Neurophysiol. 94, 429–440.

Friston, K.J., Frith, C.D., Turner, R., Frackowiak, R.S., 1995. Characterizing evoked hemodynamics with fMRI. NeuroImage 2, 157–165.

Fukushima, J., Morita, N., Fukushima, K., Chiba, T., Tanaka, S., Yamashita, I., 1990. Voluntary control of saccadic eye movements in patients with schizophrenic and affective disorders. J. Psychiatr. Res. 24, 9–24.

Gottesman, I.I., Gould, T.D., 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. Am. J. Psychiatry 160, 636–645.

Hallett, P.E., 1978. Primary and secondary saccades to goals defined by instructions. Vision Res. 18, 1279–1296.

Hariri, A.R., Weinberger, D.R., 2003. Imaging genomics. Br. Med. Bull. 65, 259–270.

Harris, M.S., Reilly, J.L., Keshavan, M.S., Sweeney, J.A., 2006. Longitudinal studies of antisaccades in antipsychotic-naive first-episode schizophrenia. Psychol. Med. 36, 485–494.

Hutton, S.B., Ettinger, U., 2006. The antisaccade task as a research tool in psychopathology: a critical review. Psychophysiology 43, 302–313.

Kimmig, H., Greenlee, M.W., Huethe, F., Mergner, T., 1999. MR-eyetracker: a new method for eye movement recording in functional magnetic resonance imaging. Exp. Brain Res. 126, 443–449.

Klein, C., Berg, P., 2001. Four-week test–retest stability of individual differences in the saccadic CNV, two saccadic task parameters, and selected neuropsychological tests. Psychophysiology 38, 704–711.

Machielsen, W.C., Rombouts, S.A., Barkhof, F., Scheltens, P., Witter, M.P., 2000. fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. 9, 156–164.

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, 955–958.

McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S., Holmes, A.P., 2000. Variability in fMRI: an examination of inter-session differences. NeuroImage 11, 708–734.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Raemaekers, M., Vink, M., van den Heuvel, M.P., Kahn, R.S., Ramsey, N.F., 2005. Brain activation related to retrosaccades in saccade experiments. NeuroReport 16, 1043–1047.

Raemaekers, M., Ramsey, N.F., Vink, M., van den Heuvel, M.P., Kahn, R.S., 2006a. Brain activation during antisaccades in unaffected relatives of schizophrenic patients. Biol. Psychiatry 59, 530–535.

Raemaekers, M., Vink, M., van den Heuvel, M.P., Kahn, R.S., Ramsey, N.F., 2006b. Effects of aging on BOLD fMRI during prosaccades and antisaccades. J. Cogn. Neurosci. 18, 594–603.

Ramsey, N.F., van den Brink, J.S., van Muiswinkel, A.M., Folkers, P.J., Moonen, C.T., Jansma, J.M., Kahn, R.S., 1998. Phase navigator correction in 3D fMRI improves detection of brain activation: quantitative assessment with a graded motor activation procedure. NeuroImage 8, 240–248.

Ramsey, N.F., Jansma, J.M., Jager, G., Van Raalten, T., Kahn, R.S., 2004. Neurophysiological factors in human information processing capacity. Brain 127, 517–525.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16, 105–113.

Roy-Byrne, P., Radant, A., Wingerson, D., Cowley, D.S., 1995. Human oculomotor function: reliability and diurnal variation. Biol. Psychiatry 38, 92–97.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59 (Suppl 20), 22–33.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing reliability. 86 ed., pp. 420-428.

Smith, S.M., Beckmann, C.F., Ramnani, N., Woolrich, M.W., Bannister, P.R., Jenkinson, M., Matthews, P.M., McGonigle, D.J., 2005. Variability in fMRI: a re-examination of inter-session differences. Hum. Brain Mapp. 24, 248–257.

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

Thomason, M.E., Foland, L.C., Glover, G.H., 2007. Calibration of BOLD fMRI using breath holding reduces group variance during a cognitive task. Hum. Brain Mapp. 28, 59–68.

van Gelderen, P., Ramsey, N.F., Liu, G., Duyn, J.H., Frank, J.A., Weinberger, D.R., Moonen, C.T., 1995. Three-dimensional functional magnetic resonance imaging of human brain on a clinical 1.5-T scanner. Proc. Natl. Acad. Sci. U. S. A. 92, 6906–6910.

Veltman, D.J., Friston, K.J., Sanders, G., Price, C.J., 2000. Regionally specific sensitivity differences in fMRI and PET: where do they come from? NeuroImage 11, 575–588.

Vink, M., Ramsey, N.F., Raemaekers, M., Kahn, R.S., 2006. Striatal dysfunction in schizophrenia and unaffected relatives. Biol. Psychiatry 60, 32–39.

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