Published in final edited form as:
Neuroimage. 2008 June ; 41(2): 311–322.

Reliability of fMRI for Studies of Language in Post-Stroke Aphasia Subjects

Kenneth P. Eaton[1], **Jerzy P. Szaflarski**[2,3,4*], **Mekibib Altaye**[1], **Angel L. Ball**[5,#], **Brett M. Kissela**[2,3], **Christi Banks**[2], and **Scott K. Holland**[1,6]

[1] *Imaging Research Center, Cincinnati Children's Hospital Research Foundation, Cincinnati, Ohio, U.S.A.*

[2] *Department of Neurology, University of Cincinnati Academic Health Center, Cincinnati, Ohio, U.S.A.*

[3] *The Neuroscience Institute, Cincinnati, Ohio, U.S.A.*

[4] *Center for Imaging Research, University of Cincinnati Academic Health Center, Cincinnati, Ohio, U.S.A.*

[5] *Department of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, Ohio, U.S.A.*

[6] *Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, U.S.A.*

Abstract

Quantifying change in brain activation patterns associated with post-stroke recovery and reorganization of language function over time requires accurate understanding of inter-scan and inter-subject variability. Here we report inter-scan variability measures for fMRI activation patterns associated with verb generation (VG) and semantic decision/tone decision (SDTD) tasks in 4 healthy controls and 4 aphasic left middle cerebral artery (LMCA) stroke subjects. A series of 10 fMRI scans was completed on a 4T Varian scanner for each task for each subject, except for one stroke subject who completed 5 and 6 scans for SDTD and VG, thus yielding 35 and 36 total stroke subject scans for SDTD and VG, respectively. Group composite and intraclass correlation coefficient (ICC) maps were computed across all subjects and trials for each task. The patterns of reliable activation for the VG and SDTD tasks correspond well to those regions typically activated by these tasks in healthy and aphasic subjects. ICCs for activation were consistently high ($R_{0.05} \approx 0.8$) for individual tasks among both control and aphasic subjects. These voxel-wise measures of reliability highlight regions of low inter-scan variability within language circuitry for control and post-recovery stroke subjects. ICCs computed from the combination of the SDTD/VG data were markedly reduced for both control and aphasic subjects as compared with the ICCs for the individual tasks. These quantitative measures of inter-scan variability support the proposed use of these fMRI paradigms for longitudinal mapping of neural reorganization of language processing following left hemispheric insult.

Keywords

stroke; language; reliability; fMRI; functional MRI; language recovery

*Correspondence to: Dr. Jerzy P. Szaflarski, University of Cincinnati Academic Health Center, Department of Neurology, Medical Sciences Building Rm. 4506, 231 Albert B. Sabin Way, Cincinnati, OH 45267-0525. E-mail: Jerzy.Szaflarski@uc.edu.
#Current address: Department of Biological and Health Sciences, Texas A&M University, Kingsville, Texas, U.S.A.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Aphasia, the loss or impairment of the ability to produce and/or comprehend language, is a common and disabling symptom of stroke. Frequently, adult patients recovering from post-stroke aphasia display redistribution of language activation patterns to the right hemisphere homotopic to the left hemispheric language areas (Weiller *et al.*, 1995; Heiss *et al.*, 1999; Thulborn *et al.*, 1999; Kim *et al.*, 2002). Although fMRI has been widely used to study the neural underpinnings of sensorimotor and neurocognitive functions in the human brain (Binder *et al.*, 1996; Cramer *et al.*, 1997; Szaflarski *et al.*, 2002; Szaflarski *et al.*, 2004; Szaflarski *et al.*, 2006a,b; Szaflarski *et al.*, 2008), its use in longitudinal studies of post-stroke brain recovery and reorganization has been limited (Fernandez *et al.*, 2004; Saur *et al.*, 2006). One hurdle to longitudinal fMRI studies of neuroplasticity is the accurate distinction between variability arising from the mechanisms of functional reorganization and confounding sources of inter-scan variability, which may arise from such factors as MRI performance, pulse sequence timing variations, subject placement, variations in physiological status, and intrinsic noise of the electronic instrumentation (Genovese *et al.*, 1997; Noll *et al.*, 1997). An accurate assessment of neuroplasticity over time requires that these inherent sources of inter-scan variability are sufficiently small. Of particular importance is the choice of an fMRI task paradigm that reliably activates the same brain regions over successive scans for a given subject.

There are a number of approaches for quantifying variability within functional data. Between-subject variability can be quantified by simply averaging activation maps across a number of subjects or employing second-level analyses like a generalized linear model (GLM) algorithm (Worsley and Friston, 1995) thus giving insight into how consistently the paradigm in question can generate activity within the same brain regions across subjects. Within-subject, or inter-scan, variability has been investigated in fMRI using scatter plots (Tegeler *et al.*, 1999; Specht *et al.*, 2003), which involves plotting for every voxel the t-value of the first measurement against that of the second one and computing the correlation coefficient of the scatter pattern. Overlap analyses, in which the overlap in volume between two thresholded parametric maps is the representative measure for reliability, has been used to quantify both within-subject and between-subject variability in fMRI activation patterns for visual processing (Rombouts *et al.*, 1998; Machielsen *et al.*, 2000; Specht *et al.*, 2003).

One method for estimating the relative contributions of within-subject and between-subject variance within a group of subjects is the intraclass correlation coefficient (ICC), a statistical analysis that has been used extensively in psychometric studies (Shrout and Fleiss, 1979; McGraw and Wong, 1996) but to a much lesser degree in fMRI studies (Specht *et al.*, 2003). Since the ICC provides for an estimate of the voxel-wise inter-scan variability of a subject group measured relative to the inter-subject variability, it is an ideal choice for quantifying the reliability of fMRI paradigms intended to be used to in the study of neuroplasticity occurring within subjects from one time point to the next. Therefore, the purpose of this study was to use the ICC to quantify the reliability of fMRI language activation paradigms intended for use in the evaluation of longitudinal changes in cortical patterns of activation following left middle cerebral artery (LMCA) stroke.

Here we use the ICC to assess reliability of the verb generation (VG) and semantic decision/tone decision (SDTD) tasks. The choice of these tasks was dictated by the fact that both tasks have been widely used in fMRI studies of language development and reorganization in response to disease or injury (Binder *et al.*, 1995, 1996, 1997; Cao *et al.*, 1999; Springer *et al.*, 1999; Holland *et al.*, 2001; Szaflarski *et al.*, 2002; Jacola *et al.*, 2006; Szaflarski *et al.*, 2006a,b; Holland *et al.*, 2007; Karunanayaka *et al.*, 2007; Tillema *et al.*, 2007; Szaflarski *et al.*, 2008). The ICC maps generated for these tasks are compared to group composite activation maps as a demonstration of which voxels activated by the tasks exhibit high reliability. Combined analyses of multiple language paradigms, such as combined task analysis (CTA) and conjunction analysis, have been shown to improve the delineation of task-independent language areas (Price *et al.*, 1997; Carpentier *et al.*, 2001; Ramsey *et al.*, 2001). As an alternative method for delineating task-independent language areas, ICC maps are generated for a combined data set that incorporates both the SDTD and VG data as if it they been generated by the same language paradigm. The findings here are intended to highlight the applicability and usefulness of ICCs in studying the neural substrates of aphasia recovery using LMCA stroke as a model for deletion of typical language areas in the brain.

Materials and Methods

Subjects

Four healthy controls and four post-stroke aphasia subjects (2–8 years after the incident stroke) were recruited for this study after signing IRB-approved consent (Table 1). If appropriate, in the case of aphasia subjects, consent was also obtained from a significant other or a caregiver. When answering questions regarding handedness (Table 1), aphasic subjects were specifically asked to remember their pre-stroke hand preference. Over the course of 10 weeks, 7 of the 8 subjects underwent five fMRI sessions during which they performed VG and SDTD tasks, each of them twice per session. One aphasic subject was only able to complete 5 scans for the SDTD task and 6 scans for the VG task. This subject developed a spontaneous subdural hematoma (diagnosed during third scanning session) and had to be withdrawn from the study. The extent of the LMCA strokes can be seen in the anatomical underlays of Fig. 3. All stroke subjects had already completed speech/language rehabilitation and their aphasia was considered to be stable. Prior to the first scanning session the healthy controls and stroke subjects received a pure tone hearing screening and a brief language evaluation battery that included subtests I, III and VI from the Revised Token Test (RTT) and the receptive language subtests (word discrimination, repetition and complex ideational) and expressive language subtests (Boston Naming Test, and responsive naming) of the Boston Diagnostic Aphasia Exam-3 (BDAE-3) Short Version. Table 2 summarizes the subjects' performance on the pre-fMRI neurolinguistic evaluations, as well as the fMRI language tasks.

Language tasks

The fMRI tasks used in this study have been utilized by our group and others for studies of language localization in healthy subjects and patients with epilepsy or stroke. A detailed description of the tasks including post-scan testing for the VG task and intra-scan testing for the SDTD task is given elsewhere (Szaflarski *et al.*, 2008). Both of these semantically-weighted language tasks have been shown to exhibit strongly left-lateralized patterns of activation in classical language areas of the inferior frontal lobe and posterior temporo-parietal region. The robust patterns of left-dominant activation and clinical utility of these tasks make them good candidates for studies of language reorganization, provided their reliability can be established. A brief description of the tasks follows.

The design of the VG blocked fMRI paradigm was based on the description by Petersen *et al*. (1988). Briefly, during the active condition, a noun was presented binaurally every 5 seconds. The subject was required to silently generate verbs associated with each noun. For example, if the noun "stove" was presented, the subject might generate the verbs "cook," "bake," or "clean." The subject was instructed to generate verbs without saying them, in order to minimize the motion artifact associated with speech. Five active conditions lasting 30 seconds each were separated by control conditions. During the six control conditions the subject was instructed to perform bilateral sequential finger tapping, starting with thumb/index opposition, in response to each FM tone centered on 400 Hz with 25% modulation presented every 5 seconds. This control condition was designed to control for the auditory prompt used in the verb generation task and to distract the subject from verb generation during rest. Excellent age-independent performance is usually seen with this verb generation task (Chiu *et al.*, 2006). Performance on this task was assessed at the conclusion of the scanning session by administering a test of recall of the nouns presented during the active condition. Performance was measured as the percentage of correct recall (Table 2).

The SDTD language paradigm was based on the task developed by Binder *et al.* (1995) and consisted of two different intervening (block design) conditions: the control condition (tone recognition; performed 8 times) and the active condition (semantic recognition; performed 7 times). Presentation of prompts for each condition lasted 30 seconds (15 seconds for the first tone recognition). During the control condition subjects were presented with brief sequences of four to seven lower- (500 Hz) and higher-pitch (750 Hz) tones. A button press response using the non-dominant hand was required for any sequence containing two higher-pitch tones. During the active condition subjects were presented with spoken English nouns designating animals and were prompted to press the “yes” button using the dominant hand when stimuli met two criteria: “native to the United States” and “commonly used by humans.” If the animal did not fulfill one or both criteria, they pressed the “no” button. Performance on this task was assessed during the scanning session by calculating the percentage of correct button responses (Table 2).

Scanning protocol

Prior to entering the scanner, all subjects learned the fMRI tasks, and their understanding of the language tasks was tested by performing a training run that included, for SDTD, a sequence of 5 sets of tones followed by a sequence of 5 nouns designating animals. For VG, the subjects listened to a sequence of 6 nouns presented every 5 seconds. Subjects were allowed to proceed to the scanner only if they responded correctly to all 10 SDTD items and if they were able to generate at least one verb associated with each of the presented nouns within the allotted time. This procedure was repeated each time the subjects returned for scanning (every 2 weeks for 10 weeks).

All scans were performed on a 4T Varian Unity INOVA Whole Body MRI/MRS scanner based on a 92.5 cm Oxford 4T actively-shielded superconducting magnet (Oxford Magnet Technology, Oxford, England). The INOVA system was controlled by a SUN workstation running Varian’s VnmrJ™ and SpinCAD™ image processing and pulse sequence development software under a Unix-based operating system. The scanner was equipped with an MRI-compatible audio/video system (Resonance Technologies Inc., Van Nuys, CA) that included gradient noise attenuating headphones. An Apple® computer was used to administer neurocognitive tasks through the audio/visual system and record subject responses.

At each scan session an initial alignment scan was performed in three orthogonal planes using fast gradient echo imaging. Necessary head position adjustments were made at this point to locate the corpus callosum and to orient the head so that the AC-PC reference line was as close as possible to the vertical axis of the scanner. Next, a high-resolution, T1-weighted 3-D MDEFT (Modified Driven Equilibrium Fourier Transform) anatomical reference brain image was obtained in the axial plane (Ugurbil *et al.*, 1993; Duewll *et al.*, 1996). Parameters for this scan were as follows: TR/TE = 13.1/6 ms, FOV = 25.6×19.2×19.2 cm, flip angle = 22°, voxel dimensions = 1×1×1 mm. Finally, functional images were obtained using a T2*-weighted, gradient-echo EPI (Echo Planar Imaging) pulse sequence with the following parameters: TR/TE = 3000/30 ms, FOV = 25.6×25.6 cm, matrix = 64×64 pixels, number of slices = 30, slice thickness = 4 mm, flip angle = 75°.

Data processing

Processing of 3-D anatomical and fMRI image data was done using Cincinnati Children's Hospital Image Processing Software (CCHIPS©) and routines written in-house using the IDL programming language (ITT Visual Information Solutions, Boulder, CO). All anatomical and functional images were transformed into the Talairach reference frame (Talairach and Tournoux, 1988) using affine spatial transformation with the brain rotated into AC-PC coordinate frame followed by linear scaling prior to statistical analyses. The statistical parametric maps for the *z* statistic were computed using the GLM (Generalized Linear Model) (Worsley and Friston, 1995) algorithm implemented within CCHIPS©. Sets of cosine basis functions were used as covariates to account for possible signal drift and aliased respiratory and cardiac signals. A second-level analysis was performed across all scans for each combination of subject group (control or aphasic) and task (VG or SDTD). The GLM algorithm was used to generate group composite activation maps which were then thresholded at $z = 1.96$ ($P_{\text{uncorrected}} = 0.05$) and limited to clusters of size greater than or equal to 5 contiguous voxels to visualize the activation patterns. A flexible factorial design analysis was performed in SPM5 (http://www.fil.ion.ucl.ac.uk/spm) to highlight differences in activation patterns between subjects and tasks. The flexible factorial analysis was chosen due to the fact that there were multiple scans performed for each subject, as opposed to a single scan for each subject in the group. The flexible factorial design analysis used a “subject by condition” design to model the interactions between subject and test condition (SDTD or VG) factors as well as modeling global effects for each subject.

Reliability measures

For the reliability measures the *z*-score activation maps were first thresholded at 1.96 ($P_{\text{uncorrected}} \leq 0.05$). This created maps of significant activation that had either a ”0” (less than the significance threshold) or a ”1” (greater than or equal to the significance threshold) for each voxel. These maps were averaged across each subject to compute the fraction of trials for which a voxel was significantly active (activation frequency). We chose to compute the intraclass correlation coefficients (ICCs) using the thresholded maps (zeroes and ones) instead of the raw *z*-score maps. Computing the reliability using the raw *z*-scores would be the equivalent of testing the voxel-wise reliability of getting the same *z*-score from scan to scan. In this case, a voxel which displayed insignificant *z*-scores from scan to scan and did so consistently (i.e., always a value of $z = 0.5$) could still have a high measure of reliability. Conversely, a voxel which always displayed significant *z*-scores but with a varying degree of significance from scan to scan (i.e., $z = 3$, 5, or 7) would have a low measure of reliability. Therefore, in order to investigate the reliability of getting significant *z*-scores (of any degree) from scan to scan, the thresholded maps were used in the reliability computations.

ICCs have been used extensively in psychometric studies (Shrout and Fleiss, 1979; McGraw and Wong, 1996) as a method for quantifying the degree of inter-rater variability relative to the degree of inter-subject variability. ICCs provide an estimate of the reliability with which multiple raters can reproduce a given diagnostic measure. In the context of the current study, the ICC is used to estimate the inter-scan (i.e., inter-rater) variability relative to the inherent variability between subjects. ICCs were computed in MATLAB (The MathWorks, Inc., Natick, MA) on a voxel-by-voxel basis across all subjects and all trials for the SDTD, VG, and combined SDTD/VG data using the following formulas (Shrout and Fleiss, 1979; McGraw and Wong, 1996):

$$R_\alpha=\frac{\mathrm{BMS}-\mathrm{WMS}}{\mathrm{BMS}+(k_0-1)\mathrm{WMS}}$$

$$\mathrm{BMS}=\frac{1}{N-1}\sum_{i=1}^{N}k_i(\overline{T}_i-\overline{T})^2 \quad \mathrm{WMS}=\frac{1}{K-N}\sum_{i=1}^{N}(k_i-1)S_i^2 \quad k_0=\overline{k}-\frac{1}{K(N-1)}\sum_{i=1}^{N}(k_i-\overline{k})^2$$

$$\overline{T}_i=\frac{1}{k_i}\sum_{j=1}^{k_i}T_{ij} \quad \overline{T}=\frac{1}{K}\sum_{i=1}^{N}k_i\overline{T}_i \quad S_i^2=\frac{1}{k_i-1}\sum_{j=1}^{k_i}(T_{ij}-\overline{T}_i)^2 \quad \overline{k}=\frac{K}{N} \quad K=\sum_{i=1}^{N}k_i$$

where, for a given voxel, R_α is the reliability as computed from data thresholded at an uncorrected *P* value of α or lower, BMS is the between-subject mean square, WMS is the within-subject mean square, T_{ij} is the activation significance (0 or 1, thresholded at $P \leq \alpha$) for subject *i* and trial *j*, T_i is the mean of T_{ij} across trials for subject *i*, T is the mean of T_{ij} across all subjects and trials, S_i is the standard deviation of T_{ij} across trials for subject *i*, k_i is the number of trials for subject *i*, K is the total number of trials across subjects, N is the number of subjects, k is the average number of trials per subject, and k_0 is the effective number of trials (necessary to compute for aphasic subjects since all values of k_i are not equal).

All ICC maps were spatially filtered using a 2 mm Gaussian filter, thresholded at R = 0.4, and limited to clusters of size greater than or equal to 10 contiguous voxels to correct for the occurrence of spurious individual voxels with high reliability. Regions of interest (ROIs) were then computed for areas of reliable activation. The following benchmarks (similar to those of Landis and Koch (1977)) were used for categorizing reliability results: 0.4–0.59 (moderate reliability), 0.6–0.79 (good reliability), 0.8–0.99 (high reliability), and 1.0 (perfect reliability). Unless otherwise noted, all significance tests were performed using a two-sample *t* test.

Results

Neurolinguistic results

The neurolinguistic baseline measures and fMRI task performance for control and aphasic subjects are summarized in Table 2. The neurolinguistic measures for control subjects formally confirmed normal language ability. Aphasic subjects tended to produce lower scores for the Word Repetition (WR) test, Boston Naming Test (BNT), and subtests III and VI of the Revised Token Test (RTT). All aphasic subjects were non-fluent and demonstrated degrees of agrammatism and reduced phrase length. Results of the BDAE-3 indicate auditory comprehension scores on word discrimination and complex ideational subtests as mild (70–80th percentile), low moderate (40–50th percentile), high moderate (60–70th percentile), and severe (20th percentile) for aphasic subjects 1 through 4, respectively. The RTT is more sensitive to short term auditory recall and temporal sequencing than the BDAE-3 comprehensive measures. Subject 1 had a behavior of verbalizing instructions in a self-talk approach, a strategy which affected his score but resulted in accurate responses. Subject 2 had the most difficulty retaining information on subtest VI, and seemed to perseverate on touching rather than picking up the item even with repetition of directions. Subject 3 demonstrated mild frustration and produced decreasing scores as the task difficulty increased. Subject 4 appeared to benefit from task practice, scoring higher as the task difficulty increased.

FMRI behavioral results

For the VG task, both control and aphasic subjects produced consistent post-scan noun recall results (see Table 2) with slightly lower scores for aphasic subjects relative to control subjects. Aphasic subjects also produced lower scores than control subjects for both the semantic and tone decision portions of the SDTD task. Differences in task performance were also seen when the aphasic subject population was subdivided into those with primarily inferior frontal lesions (aphasic subjects 1 and 3) and those with primarily temporo-parietal lesions (aphasic subjects 2 and 4). Although there was little difference between the two subpopulations for the VG task performance, subjects with inferior frontal lesions tended to perform better on the SDTD task than subjects with temporo-parietal lesions.

Group composites

Group activation maps for control subjects (Fig. 1, top) correspond well to patterns typically seen for the VG (Cao *et al.*, 1999;Holland *et al.*, 2001;Jacola *et al.*, 2006;Szaflarski *et al.*, 2006a,b;Holland *et al.*, 2007) and SDTD (Binder *et al.*, 1995,1996,1997;Springer *et al.*, 1999;Szaflarski *et al.*, 2002;Szaflarski *et al.*, 2008) tasks. The SDTD task strongly activates left frontal gyri (superior, medial, and inferior), left temporal gyri (medial and superior), bilateral cingulate gyri, and precuneus. The VG task primarily activates left inferior and medial frontal gyri. Factorial analysis (data not shown) indicates that there is significantly larger ($P_{\text{corrected}} < 0.05$) activation predominantly in left temporal gyri, bilateral cingulate gyri, and precuneus for the SDTD task relative to the VG task, while the VG task generates slightly larger activation in the posterior portion of inferior frontal gyrus than the SDTD task.

Group activation maps for aphasic subjects (Fig. 1, bottom) show similar overall patterns to control subjects, albeit with greatly reduced strength of activation and slightly more bilateral or right-hemispheric activation. This atypical pattern is not unexpected as it has been seen in patients with history of LMCA stroke (Jacola *et al.*, 2006) or epilepsy (Springer *et al.*, 1999;Yuan *et al.*, 2006;Szaflarski *et al.*, 2008). Activation is seen in left inferior frontal gyrus, bilateral medial and superior frontal gyri, bilateral angular gyri, bilateral cingulate gyri, and precuneus for the SDTD task. The VG task activates left medial frontal gyrus, bilateral inferior frontal gyri, right superior temporal gyrus, and bilateral angular gyri. For aphasic subjects, factorial analysis indicates significantly larger activation in right angular gyrus, bilateral cingulate gyri, and precuneus for the SDTD task relative to the VG task, while the VG task generates slightly larger activation in right superior temporal gyrus and the posterior portion of medial frontal gyrus compared to the SDTD task.

Factorial analysis was also performed for comparisons between control and aphasic subject groups for both tasks. For the SDTD task, control subjects displayed larger activation in left medial frontal gyrus, while aphasic subjects displayed larger activation within portions of bilateral superior temporal gyri. For the VG task, control subjects appeared to have no activation that was significantly greater than aphasic subjects, while aphasic subjects displayed slightly larger activation than controls occurring diffusely throughout right inferior and medial frontal gyri, bilateral medial and superior temporal gyri, and cingulate gyrus. The minor differences in activation between subject groups for the VG task appear to mirror the behavioral data (Table 2), in which VG performance is not significantly different between control and aphasic subjects.

Reliability measures

Functional MRI reliability results for control and aphasic subjects are shown in Fig. 2 and Fig. 3, respectively, for selected axial slices. The fraction of trials for which a voxel activates (activation frequency) for each of the four control/aphasic subjects is shown in the top four rows (labeled “1” through “4”) for the SDTD task (left column) and VG task (right column). The reliability computed across all scans for all four subjects is shown in the bottom row (labeled “R”) for each task, and the lower inset shows the reliability computed by combining the SDTD and VG scans for each subject. The reliable regions of interest found for control and aphasic subjects are listed in Table 3 and Table 4, respectively. For each ROI the anatomical brain regions, maximum reliability, number of voxels, and Talairach coordinates of the ROI centroid are listed.

All of the ROIs for reliable SDTD activation in control subjects listed in Table 3 contain voxels with moderate to good reliability, with high reliability seen in left medial/superior temporal gyrus, left supramarginal gyrus, left angular gyrus, and bilateral cingulate gyrus and precuneus. Voxels with perfect reliability were also seen bilaterally in medial/superior frontal gyrus. Although fewer ROIs are seen for the VG task relative to the SDTD task, the degree of reliability contained within them is still comparable. All ROIs contain voxels with moderate to good reliability, with high reliability punctuating right inferior frontal gyrus and left inferior parietal lobe and voxels with perfect reliability seen in left medial frontal gyrus and left precentral gyrus.

Aphasic subjects display similar reliability maps (Fig. 3) to those of control subjects (Fig. 2). Although fewer ROIs are seen for the SDTD task for aphasic subjects (Table 4) relative to control subjects, the average maximum reliability across ROIs is not significantly different ($P = 0.43$). High reliability is seen in right inferior temporal gyrus, right inferior frontal gyrus, left superior temporal gyrus, and bilateral cingulate gyrus and precuneus, while voxels with perfect reliability appear in right medial frontal gyrus, left superior temporal gyrus, and left angular gyrus. As with control subjects, fewer ROIs are seen for the VG task relative to the SDTD task for aphasic subjects. All ROIs contain voxels with moderate to good reliability, with right inferior frontal gyrus containing voxels with perfect reliability.

Group composite versus ICC

Figure 4 illustrates how the GLM group composite calculation and the ICC measure differ. Three voxels (A, B, and C) within a single slice from the control subject SDTD results are selected as examples and these are highlighted in both the group composite (Fig. 4, top left) and ICC (Fig. 4, bottom left) maps. Voxel A has a low *z*-score ($z = 1.49$) but perfect reliability ($R_{0.05} = 1.00$). Voxel B represents the other extreme, having a high *z*-score ($z = 7.52$) but poor reliability ($R_{0.05} = 0.00$). Voxel C has both a high *z*-score ($z = 5.96$) and a good reliability ($R_{0.05} = 0.635$). The graphs on the right of Fig. 4 show in red the *z*-score for the given voxel across all 40 trials. The dotted red line indicates the *z*-score threshold of $z = 1.96$ ($\alpha = 0.05$) and the blue line indicates trials in which significant activation, above this threshold (indicated by a value of “1”), occurs for the selected voxel. The vertical dashed lines separate the 40 trials into groups of 10 each, one group for each subject (1 through 4, from left to right). For voxel A (Fig. 4A, top right), subjects 1, 2, and 4 have *z*-scores that never exceed 1.96, while subject 3 has *z*-scores that always exceed the $z = 1.96$ threshold, yielding no inter-scan variability within each subject. The inter-scan reliability is therefore perfect. For voxel B (Fig. 4B, middle right), the blue line illustrates how the *z*-score for each subject tends to fluctuate around the threshold of $z = 1.96$ across scans for each subject. The percentage of trials having significant activation for subjects 1 through 4 is 50%, 80%, 80%, and 70%, respectively, which results in poor inter-scan reliability. Finally, voxel C (Fig. 4C, lower right) illustrates that subject 1 never has significant *z*-scores (no inter-scan variability), subject 2 always has significant *z*-scores (again, no inter-scan variability), and subjects 3 and 4 have mostly significant *z*-scores (80% and 70%, respectively). The resulting reliability for voxel C is 0.635, in between the values for voxels A and B, due to the mixture of inter-scan and inter-subject variability.

The dependence of the ICC on the choice of the threshold selected for activation is illustrated in Figure 4D. ICC maps were generated from the control subject SDTD data at values of α (threshold for significant activation) that varied from 0.001 to 0.8 (*z* thresholds of 3.29 to 0.253, respectively). Since many voxels had little or no activity, such as those outside the brain, a subset of voxels was chosen that had an $R_\alpha > 0.4$ for at least one of these α values. The R_α values for these voxels were sorted in ascending order and plotted as a 2D histogram of the cumulative number of voxels with a given reliability (R_α shown in color) as a function of the α threshold. Figure 4D displays this histogram, illustrating the total number of voxels with a given reliability as a function of α. The color range is indicated at right, with undefined values (indicating that a voxel is never active at that α) shown in white. The dashed line represents $R_\alpha = 0.4$. The total numbers of voxels with a reliability in the good to high range ($0.6 \leq R_\alpha \leq 0.99$) remain relatively stable, while the numbers of voxels with moderate to poor reliability ($R_\alpha < 0.6$) vary a great deal with respect to α. Moving the threshold from $\alpha = 0.05$ to $\alpha = 0.02$ results in an average drop of 0.042 in the reliability (mean of the changes in R_α across all voxels). Moving the threshold from $\alpha = 0.05$ to $\alpha = 0.01$ or $\alpha = 0.001$ results in average drops in reliability of 0.071 and 0.16, respectively.

Combined tasks

Combining SDTD and VG tasks results in substantially fewer reliable voxels with lower average reliabilities compared to each task individually. For control subjects (Table 3 and lower inset of Fig. 2), moderate reliability is seen in left inferior/medial frontal gyrus and good reliability seen in left medial/superior frontal gyrus, left medial temporal gyrus, and left angular gyrus. Aphasic subjects (Table 4 and lower inset of Fig. 3) display good reliability in left and right medial frontal gyri. Taking a sample across all ROIs of the 30 voxels with the largest reliabilities, the mean of the combined SDTD/VG voxels is significantly less for both control subjects ($P < 0.001$ for combined versus either SDTD or VG) and aphasic subjects ($P < 0.001$ for combined versus either SDTD or VG).

Discussion

The purpose of this study was to quantify the reliability of two fMRI paradigms (VG and SDTD) for monitoring longitudinal changes in cortical patterns of activation during post-stroke recovery. High measures of reliability are seen for these two paradigms in regions typical for language processing in healthy subjects as well as contralateral cortical language homologues in aphasic subjects. The reliability decreases substantially when activation patterns for the two paradigms are combined suggesting that those cortical regions reliably activated by each task have little overlap: not surprising given the fact that different components of language processing are involved in verb generation versus semantic decision. These findings provide a basis for better quantifying longitudinal variability in fMRI language activation occurring during recovery from aphasia.

Reliability measures

For the current study, an ICC ($R_{0.05}$) of 0.8 (high) corresponds to a between-subject mean square (BMS) that is approximately 35 or 41 times larger than the within-subject mean square (WMS) for aphasic or control subjects, respectively. The scale factor drops to approximately 14 and 16, respectively, when $R_{0.05} = 0.6$ (good). Therefore, in the ROIs that demonstrate good or greater reliability, much of the variability in voxel activation is attributable to inter-subject sources as opposed to intra-subject (i.e., inter-scan) sources. From scan to scan, these regions display consistent activation/non-activation for a given subject. This finding supports the use of the VG and SDTD fMRI tasks in longitudinal studies of stroke patients during recovery from aphasia as a reliable means to study neuroplasticity. Changes in brain activation occurring across scans in an individual can be quantified in terms of significance relative to the measured reliability in specific ROIs using the reliability measure presented here.

Regions of greatest reliability found here correspond well to those regions most typically activated by the VG (Cao *et al.*, 1999; Holland *et al.*, 2001; Jacola *et al.*, 2006; Szaflarski *et al.*, 2006a,b) and SDTD (Binder *et al.*, 1995, 1996, 1997; Springer *et al.*, 1999; Szaflarski *et al.*, 2002) tasks (Fig. 1). VG activation in control subjects displayed good reliability in left inferior frontal gyrus, high reliability in right inferior frontal gyrus and left inferior parietal lobe, and perfect reliability in left medial frontal and precentral gyri. VG activation in aphasic subjects displayed good reliability in left medial temporal and bilateral frontal gyri and perfect reliability in right inferior frontal gyrus. Reliability of SDTD activation in control subjects was good in left and right inferior frontal gyri and right medial/superior temporal gyrus, high in left medial/superior temporal gyrus, precuneus, and angular, cingulate, and supramarginal gyri, and perfect in medial/superior frontal gyrus. Finally, reliability of SDTD activation in aphasic subjects was good in left inferior frontal gyrus, high in right inferior frontal gyrus, right inferior temporal gyrus, precuneus, and cingulate gyrus, and perfect in right medial frontal, left superior temporal, and left angular gyri.

The greater reliability measured in right versus left inferior frontal gyri for aphasic subjects supports the use of the current paradigms for studying language reorganization, because we expect subjects with early insult are more likely to demonstrate increased right hemispheric participation (Müller *et al.*, 1999; Cao *et al.*, 1999). Since this is one of the regions to which language processing is likely to migrate after a left hemispheric stroke (Thulborn *et al.*, 1999; Jacola *et al.*, 2006; Saur *et al.*, 2006), a high measure of inter-scan reliability here means that time-dependent neuroplasticity can be better discerned. Greater reliability in right inferior frontal gyrus in control subjects may be explained by the more focal nature of activation in contralateral homologues as compared with a tendency for more diffuse activation patterns in the left hemisphere (note SDTD activation frequency data for control subjects, left column of Fig. 2). In addition, this activation of the contralateral homologue only appears repeatedly in subject 3, a situation similar to that of Fig. 4A in that it would tend to generate high reliability values but would be unlikely to show up significantly on a group composite. This may suggest the presence of a smaller, more focally activated language network in the non-dominant right hemisphere in certain subjects.

The ROIs with reliable activation in control (Table 3) and aphasic (Table 4) subjects encompass many of the same brain regions, with some exceptions. Control subjects display regions of reliable activation that are no longer present in post-recovery aphasic subjects, such as the medial temporal gyri, supramarginal gyri, and paracentral and inferior parietal lobes in the SDTD task and the angular, precentral, and medial occipital gyri and precuneus in the VG task. Of particular interest are the regions of reliable activation present in aphasic subjects that are not seen in control subjects, such as the left parahippocampal and right inferior temporal gyri in the SDTD task and the right medial and bilateral superior frontal gyri in the VG task. To accurately quantify reorganization of language to these areas requires knowledge of the baseline variability in these areas in controls (i.e., pre-stroke condition). The control data can therefore be reexamined to determine if these areas display either unreliable activation or reliable non-activation (i.e., consistently no activation present from trial to trial across all subjects, which can generate undefined values of the ICC that would not be included in the data of Table 3).

When the left parahippocampal gyrus ROI for the SDTD task in the aphasic subjects is applied to the control subject reliability map, 48.4 % (15 of 31) of the voxels in the ROI have perfectly reliable non-activation (i.e., never display activation on any trial for any subject). For the right inferior temporal gyrus ROI, 66.7 % (8 of 12) of the control subject voxels have perfectly reliable non-activation. When the bilateral superior frontal gyrus ROI for the VG task in the aphasic subjects is applied to controls, 50 % (5 of 10) of the voxels have perfectly reliable non-activation. The right medial frontal gyrus ROI in the aphasic subjects overlaps slightly with the left medial frontal gyrus ROI in the control subjects for the VG task, since these ROIs are so close to the midline. These two ROIs can be considered equivalent, encompassing slightly bilateral activation of the medial frontal gyrus occurring with good reliability in both control and aphasic subjects. Overall, these brain regions display reliable activation/non-activation in both control and aphasic subjects, which is important for accurately mapping neuroplasticity in these areas during reorganization.

Group composite versus ICC

The comparison of a second-level group composite and an ICC displayed in Fig. 4 illustrates how the estimates of reliability used herein are dependent upon the inter-scan variability within each subject. This comparison also highlights ways in which the voxel-wise reliability may vary as a function of the threshold initially used to generate the maps of significant activation. A liberal threshold of $z = 1.96$ ($\alpha = 0.05$) was chosen here to allow for the inclusion of more significantly-activated voxels in the analysis. Figure 4D illustrates that the total number of voxels displaying high reliability ($R_\alpha \geq 0.8$) would remain relatively stable when moving to stricter α thresholds. The number of voxels with good reliability (R_α = 0.6—0.79) drops slightly, but voxels in this range of reliability are still relatively insensitive to the α threshold. Conversely, the number of voxels with moderate reliability (R_α = 0.4—0.59) varies more dramatically with the α threshold. The number of voxels with poor reliabilities ($R_\alpha< 0.4$) increases with stricter α, as well as the number of voxels with undefined reliabilities (i.e. the threshold is so high that there are many more voxels that never display suprathreshold activation).

The dependence of the ICC on the significance threshold, α, is further illustrated by the graphs in Figure 4A—C showing three selected voxels that exhibit different reliabilities. For voxel A (Fig. 4A), the reliability can only decrease as the threshold is raised. Nothing will change for subjects 1, 2, or 4, but eventually the inter-scan variability for subject 3 will increase with increasing values for the significance threshold, causing an overall decrease in reliability at this voxel. For voxel B (Fig. 4B), an increase of the significance threshold into the range of 2.0 to 4.0 will cause a further drop of the reliability since the *z*-score (red line) will be fluctuating around the threshold, producing high within-subject variability for subjects 1, 3, and 4. The reliability will transiently increase, but still be relatively low, when the threshold increases above 4.0 and the number of significant activations (blue line) for subjects 1, 3, and 4 drops towards zero (i.e., consistent non-activation). Voxel C (Fig. 4C) will show transient increases in reliability as the significance threshold increases, mainly due to the number of significant activations for subject 4 dropping to zero at $z > 3.4$.

Combined task results

ICCs computed for the combined SDTD/VG data (insets of Fig. 2 and Fig. 3) are markedly reduced for both control and aphasic subjects as compared with the ICCs for the individual tasks. This is likely due to non-intersecting patterns of activation for the two tasks. Although both paradigms are designed to engage left-dominant language circuitry involved in verbally non-productive tasks (covert generation for VG and a button press response for SDTD), they are often noted to differentially activate both classical and non-classical cortical language regions. Both are shown to elicit activation around Broca's area and occasionally the right hemisphere homologue (Holland *et al.*, 2001;Binder *et al.*, 1997). Accordingly, the combined SDTD/VG results for control subjects (Table 3) display regions of good reliability in left inferior, medial, and superior frontal gyri, but no reliable right hemispheric activation. However, combined results for aphasic subjects (Table 4) show regions of good reliability in both left and right medial frontal gyrus, as reorganization of language to the right hemisphere tends to increase the reliability of contralateral cortical activation.

With regard to Wernicke's region, the VG paradigm typically activates left superior and medial temporal gyri (Holland *et al.*, 2001), while activity with the SDTD paradigm is often seen in left inferior and medial temporal gyri (Binder *et al.*, 1997). The overlap between these two, the left medial temporal gyrus, is shown to activate with good reliability in the combined data for control subjects (Table 3). The combined data for aphasic subjects (Table 4) does not include reliable activation around Wernicke's, which may be due to the heterogeneity in LMCA lesions between subjects. Subdividing the aphasic subjects based on lesion site may generate reliability maps for the SDTD data, and thus the combined SDTD/VG data, with higher average reliabilities, but the subject population used here is not large enough to investigate this.

Limitations

The current study focused on quantifying the ratio of inter-scan and inter-subject variability for the given task paradigms on a voxel-by-voxel basis. With a sufficient number of repeat scans per subject we have been able to make a good estimate of the inter-scan variability. However, the small subject population (N = 4 for each, control and aphasic subjects) introduces some limitations. First, we may not be capturing the full spectrum of inter-subject variability. This would limit the applicability of the current findings in a larger subject population. Second, the small aphasic subject population means that heterogeneity in lesion properties may contribute to significant variability in activation patterns. Finally, the small sample size also limits the investigation of features among subpopulations of the subjects. Even so, the group composite maps for control and aphasic subjects (Fig. 1) correspond well to the typical patterns of activation seen for the VG and SDTD tasks in healthy subjects, suggesting that the sample is representative as a basis for the analyses intended in this study. The effect of increasing the number of subjects on the reliability measures depends on the relative contributions additional subjects will make to the between-subject variability (BMS) and the within-subject variability (WMS). If additional subjects have consistent inter-scan activity, but differ substantially from one another (high inter-subject variability), increases in the reliability measures would be expected. Overall, this preliminary data helps to support the feasibility of using the VG and SDTD paradigms in the study of longitudinal stroke recovery.

In conclusion, this study shows excellent reliability of fMRI language activation patterns in healthy and aphasic subjects when the VG and SDTD tasks are considered separately. The reliability of activation for the combined set of tasks is decreased compared to individual tasks, which may reflect differences in task design or involvement of different structures underlying language production in the SDTD versus VG tasks. These quantitative measures of variability support the proposed use of these two fMRI paradigms for the longitudinal mapping of neural reorganization of language processing following left hemispheric insult.

Acknowledgments

This study was supported by a grant from The Neuroscience Institute in Cincinnati, OH, USA (JPS). Dr. Eaton is supported by a training grant from the NIH (T32-EB01656). This study was presented in part at the 59th Annual Meeting of the American Academy of Neurology (Boston, MA, 5/07) and in part at the 13th Meeting of the Organization for HBM (Chicago, IL, 6/07).

References

Binder JR, Rao SM, Hammeke TA, Frost JA, Bandettini PA, Jesmanowicz A, Hyde JS. Lateralized human brain language systems demonstrated by task subtraction functional magnetic resonance imaging. Arch. Neurol 1995;52:593–601. [PubMed: 7763208]

Binder JR, Swanson SJ, Hammeke TA, Morris GL, Mueller WM, Fischer M, Benbadis S, Frost JA, Rao SM, Haughton VM. Determination of language dominance using functional MRI: a comparison with the Wada test. Neurology 1996;46:978–984. [PubMed: 8780076]

Binder JR, Frost JA, Hammeke TA, Cox RW, Rao SM, Prieto T. Human brain language areas identified by functional magnetic resonance imaging. J. Neurosci 1997;17:353–362. [PubMed: 8987760]

Cao Y, Vikingstad EM, George KP, Johnson AF, Welch KMA. Cortical language activation in stroke patients recovering from aphasia with functional MRI. Stroke 1999;30:2331–2340. [PubMed: 10548667]

Carpentier A, Pugh KR, Westerveld M, Studholme C, Skrinjar O, Thompson JL, Spencer DD, Constable RT. Functional MRI of language processing: Dependence on input modality and temporal lobe epilepsy. Epilepsia 2001;42:1241–1254. [PubMed: 11737158]

Chiu C-YP, Schmithorst VJ, Brown RD, Holland SK, Dunn S. Making memories: A cross-sectional investigation of episodic memory encoding in children using fMRI. Dev. Neuropsych 2006;29:321–340.

Cramer SC, Nelles G, Benson RR, Kaplan JD, Parker RA, Kwong KK, Kennedy DN, Finklestein SP, Rosen BR. A functional MRI study of subjects recovered from hemiparetic stroke. Stroke 1997;28:2518–2527. [PubMed: 9412643]

Duewll S, Wolff SD, Wen H, Balaban RS, Jezzard P. MR imaging contrast in human brain tissue: Assessment and optimization at 4 T. Radiology 1996;199:780–786. [PubMed: 8638005]

Fernandez B, Cardebat D, Demonet J-F, Joseph PA, Mazaux J-M, Barat M, Allard M. Functional MRI follow-up study of language processes in healthy subjects and during recovery in a case of aphasia. Stroke 2004;35:2171–2176. [PubMed: 15297629]

Genovese CR, Noll DC, Eddy WF. Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. Magn. Reson. Med 1997;38:497–507. [PubMed: 9339452]

Heiss W-D, Kessler J, Thiel A, Ghaemi M, Karbe H. Differential capacity of left and right hemispheric areas for compensation of poststroke aphasia. Ann. Neurol 1999;45:430–438. [PubMed: 10211466]

Holland SK, Plante E, Byars AW, Strawsburg RH, Schmithorst VJ, Ball WS Jr. Normal fMRI brain activation patterns in children performing a verb generation task. NeuroImage 2001;14:837–843. [PubMed: 11554802]

Holland SK, Vannest J, Mecoli M, Jacola LM, Tillema JM, Karunanayaka PR, Schmithorst VJ, Yuan W, Plante E, Byars AW. Functional MRI of language lateralization during development in children. Int. J. Audiol 2007;46:533–551. [PubMed: 17828669]

Jacola LM, Schapiro MB, Schmithorst VJ, Byars AW, Strawsburg RH, Szaflarski JP, Plante E, Holland SK. Functional magnetic resonance imaging reveals atypical language organization in children following perinatal left middle cerebral artery stroke. Neuropediatrics 2006;37:46–52. [PubMed: 16541368]

Karunanayaka PR, Holland SK, Yuan W, Altaye M, Jones BV, Michaud LJ, Walz NC, Wade SL. Neural substrate differences in language networks and associated language behavioral impairments in children with TBI: A preliminary fMRI investigation. NeuroRehab 2007;22:1–15.

Kim YH, Ko MH, Parrish TB, Kim HG. Reorganization of cortical language areas in patients with aphasia: a functional MRI study. Yonsei Med. J 2002;43:441–445. [PubMed: 12205731]

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]

Machielsen WC, Rombouts SA, Barkhof F, Scheltens P, Witter MP. fMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp 2000;9:156–164. [PubMed: 10739366]

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psych. Meth 1996;1:30–46.

Müller R-A, Rothermel RD, Behen ME, Muzik O, Chakraborty PK, Chugani HT. Language organization in patients with early and late left-hemisphere lesion: a PET study. Neuropsychologia 1999;37:545–557. [PubMed: 10340314]

Noll DC, Genovese CR, Nystrom LE, Vazquez AL, Forman SD, Eddy WF, Cohen JD. Estimating test-retest reliability in functional MR imaging. II: Application to motor and cognitive activation studies. Magn. Reson. Med 1997;38:508–517. [PubMed: 9339453]

Petersen SE, Fox PT, Posner MI, Mintum M, Raichle ME. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 1988;331:585–589. [PubMed: 3277066]

Price CJ, Moore CJ, Friston KJ. Subtractions, conjunctions, and interactions in experimental design of activation studies. Hum. Brain Mapp 1997;5:264–272.

Ramsey NF, Sommer IEC, Rutten GJ, Kahn RS. Combined analysis of language tasks in fMRI improves assessment of hemispheric dominance for language functions in individual subjects. NeuroImage 2001;13:719–733. [PubMed: 11305899]

Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Scheltens P. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 1998;16:105–113. [PubMed: 9508267]

Saur D, Lange R, Baumgaertner A, Schraknepper V, Willmes K, Rijntjes M, Weiller C. Dynamics of language reorganization after stroke. Brain 2006;129:1371–1384. [PubMed: 16638796]

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psych. Bull 1979;86:420–428.

Specht K, Willmes K, Shah NJ, Jäncke L. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 2003;17:463–471. [PubMed: 12655586]

Springer JA, Binder JR, Hammeke TA, Swanson SJ, Frost JA, Bellgowan PSF, Brewer CC, Perry HM, Morris GL, Mueller WM. Language dominance in neurologically normal and epilepsy subjects: A functional MRI study. Brain 1999;122:2033–2046. [PubMed: 10545389]

Szaflarski JP, Binder JR, Possing ET, McKiernan KA, Ward BD, Hammeke TA. Language lateralization in left-handed and ambidextrous people: fMRI data. Neurology 2002;59:238–244. [PubMed: 12136064]

Szaflarski JP, Holland SK, Schmithorst VJ, Dunn RS, Privitera MD. High-resolution functional MRI at 3T in healthy and epilepsy subjects: hippocampal activation with picture encoding task. Epilepsy Behav 2004;5:244–252. [PubMed: 15123027]

Szaflarski JP, Holland SK, Schmithorst VJ, Byars AW. fMRI study of language lateralization in children and adults. Hum. Brain Mapp 2006a;27:202–212. [PubMed: 16035047]

Szaflarski JP, Schmithorst VJ, Altaye M, Byars AW, Ret J, Plante E, Holland SK. A longitudinal functional magnetic resonance imaging study of language development in children 5 to 11 years old. Ann. Neurol 2006b;59:796–807. [PubMed: 16498622]

Szaflarski JP, Holland SK, Jacola LM, Lindsell C, Privitera MD, Szaflarski M. Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy. Epilepsy Behav 2008;12:74–83. [PubMed: 17964221]

Talairach, J.; Tournoux, P. Co-planar stereotaxic atlas of the human brain. New York: Thieme Medical Publishers, Inc.; 1988.

Tegeler C, Strother SC, Anderson JR, Kim SG. Reproducibility of BOLD-based functional MRI obtained at 4 T. Hum. Brain Mapp 1999;7:267–283. [PubMed: 10408770]

Thulborn KR, Carpenter PA, Just MA. Plasticity of language-related brain function during recovery from stroke. Stroke 1999;30:749–754. [PubMed: 10187873]

Tillema J-M, Byars AW, Jacola LM, Schapiro MB, Schmithorst VJ, Szaflarski JP, Holland SK. Cortical reorganization of language functioning following perinatal left MCA stroke. Brain Lang. 2007Epub ahead of print

Ugurbil K, Garwood M, Ellermann J, Hendrich K, Hinke R, Hu X, Kim SG, Menon R, Merkle H, Ogawa S, Salmi R. Imaging at high magnetic fields: Initial experience at 4T. Magn. Reson. Quart 1993;9:259–277.

Weiller C, Isensee C, Rijntjes M, Huber W, Muller S, Bier D, Dutschka K, Woods RP, Noth J, Diener HC. Recovery from Wernicke's aphasia: a positron emission tomographic study. Ann. Neurol 1995;37:723–732. [PubMed: 7778845]

Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited—again. NeuroImage 1995;2:173–181. [PubMed: 9343600]

Yuan W, Szaflarski JP, Schmithorst VJ, Schapiro M, Byars AW, Strawsburg RH, Holland SK. fMRI shows atypical language lateralization in pediatric epilepsy patients. Epilepsia 2006;47:593–600. [PubMed: 16529628]

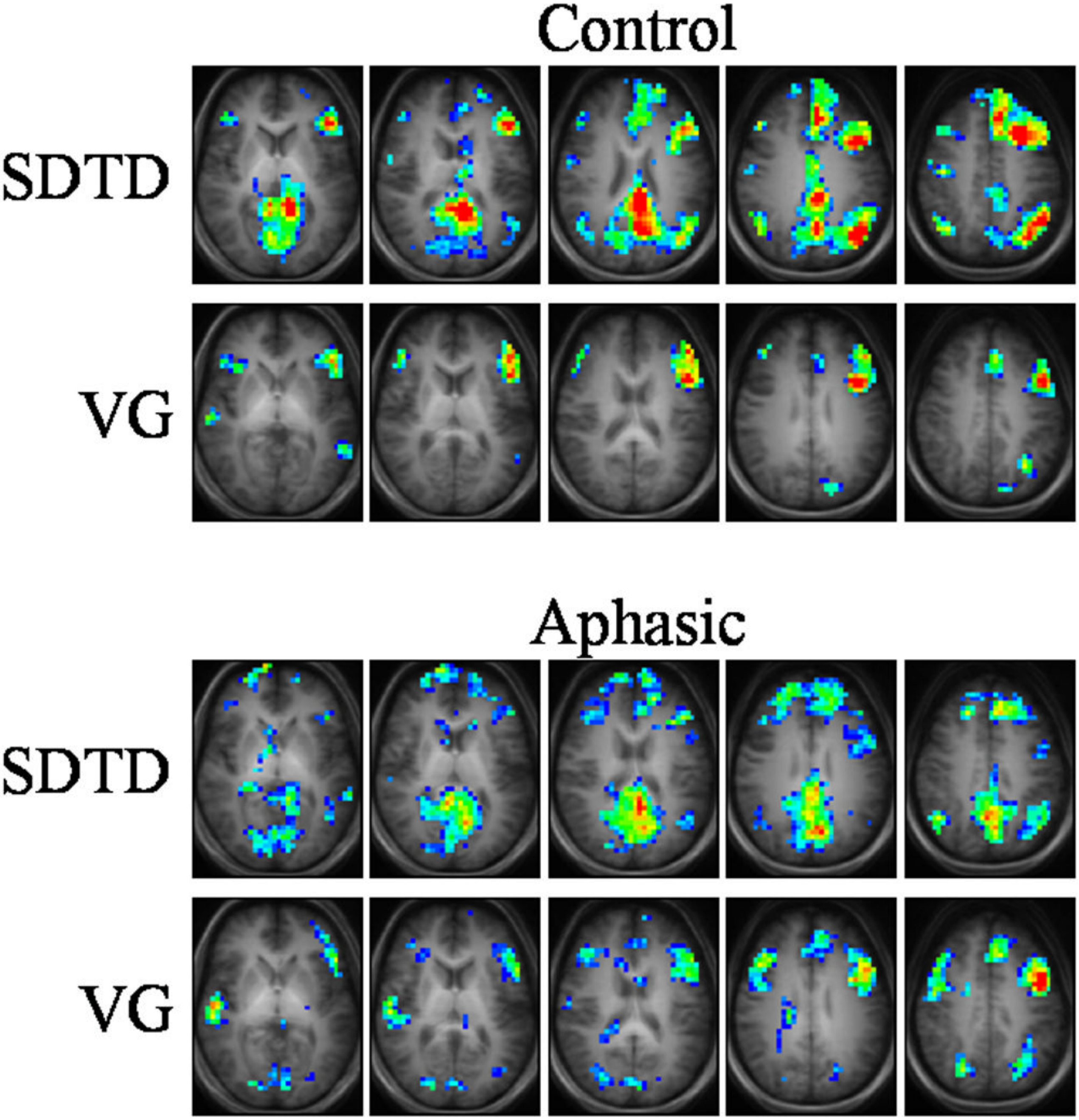

Figure 1.

Group composite maps for control (top) and aphasic (bottom) subjects. Anatomical template for control subjects used for underlays. Color scale: $z = 1.96$ (dark blue), $z \geq 6.0$ (red). Talairach z-coordinates: +7 to +39 (left to right) for control SDTD, +3 to +35 (left to right) for all others.

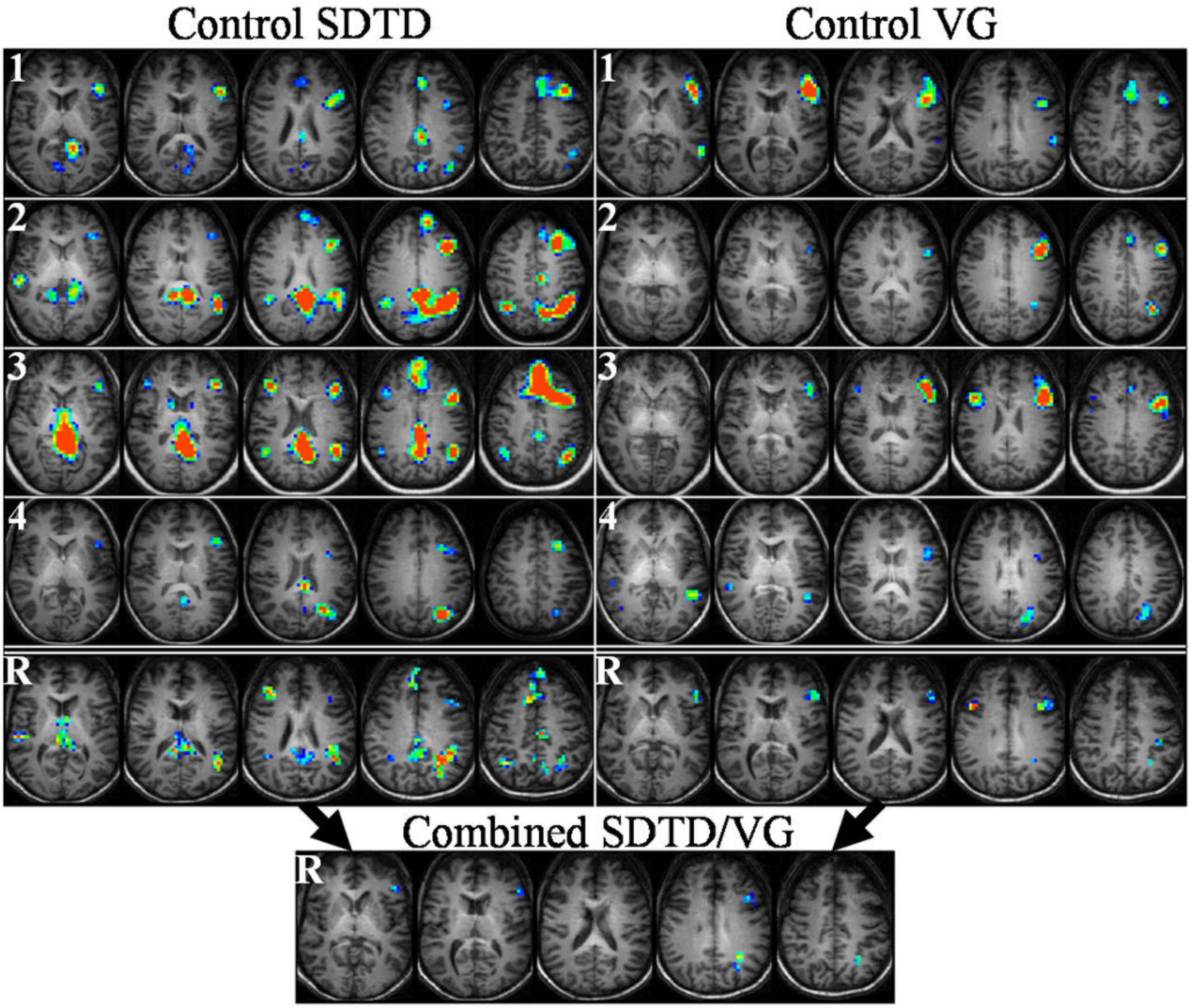

Figure 2.
Control subject activation during SDTD (left) and VG (right) tasks. Top four rows (1–4) show voxel activation frequencies for each subject, while bottom row (R) shows ICCs for each task (subject 1 used for anatomical underlay). Inset at bottom shows reliability when SDTD and VG tasks are combined for each subject. Color range: dark blue (0.4), red (0.7–1.0). Talairach z-coordinates: +7 to +39 (left to right) for SDTD, +3 to +35 (left to right) for VG and SDTD/VG.

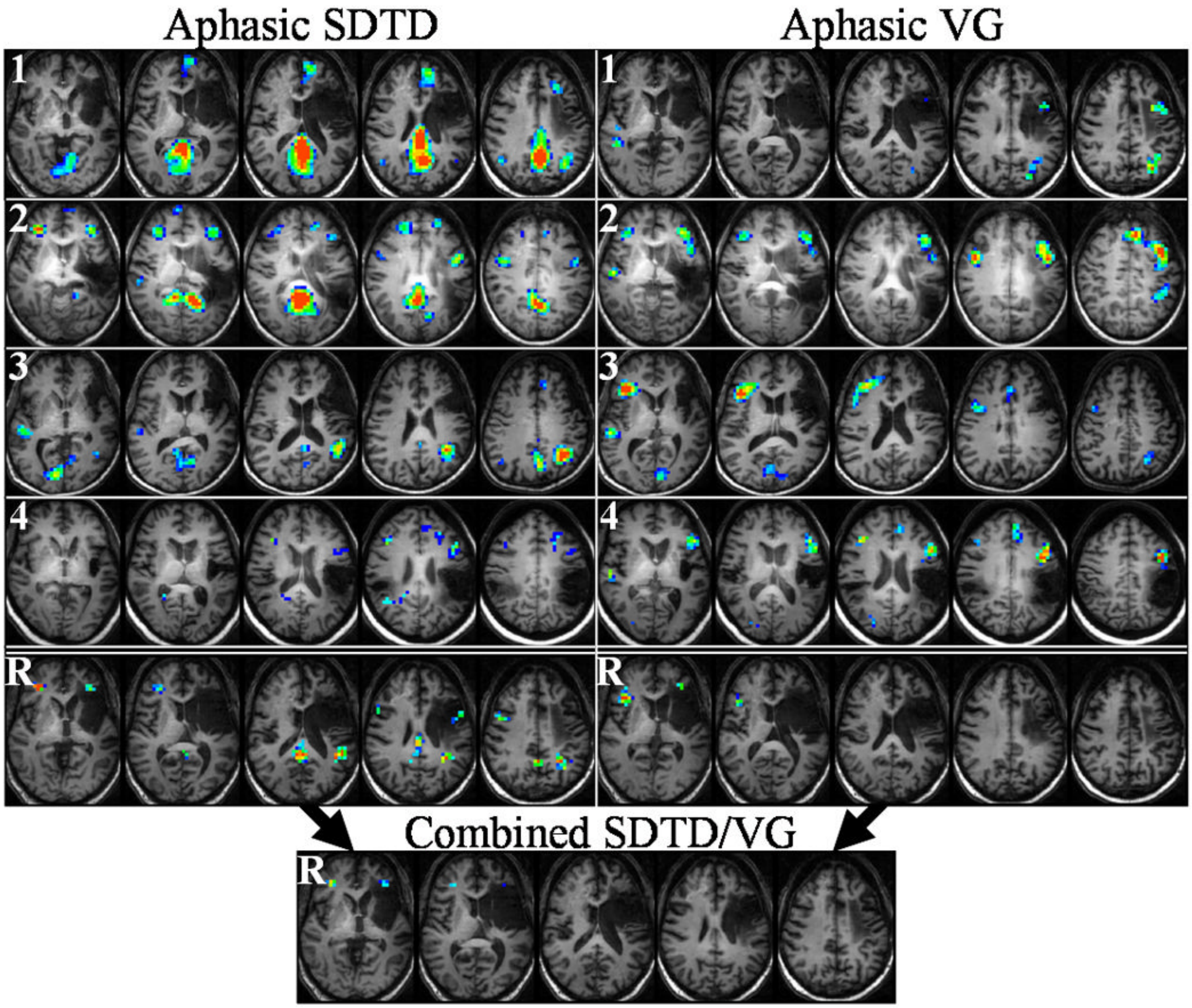

Figure 3.

Aphasic subject activation during SDTD (left) and VG (right) tasks. Top four rows (1–4) show voxel activation frequencies measures for each subject, while bottom row (R) shows ICCs for each task (subject 1 used for anatomical underlay). Inset at bottom shows reliability when SDTD and VG tasks are combined for each subject. Color range: dark blue (0.4), red (0.7–1.0). Talairach z-coordinates: +1 to +33 (left to right) for SDTD and SDTD/VG, +5 to +37 (left to right) for VG.

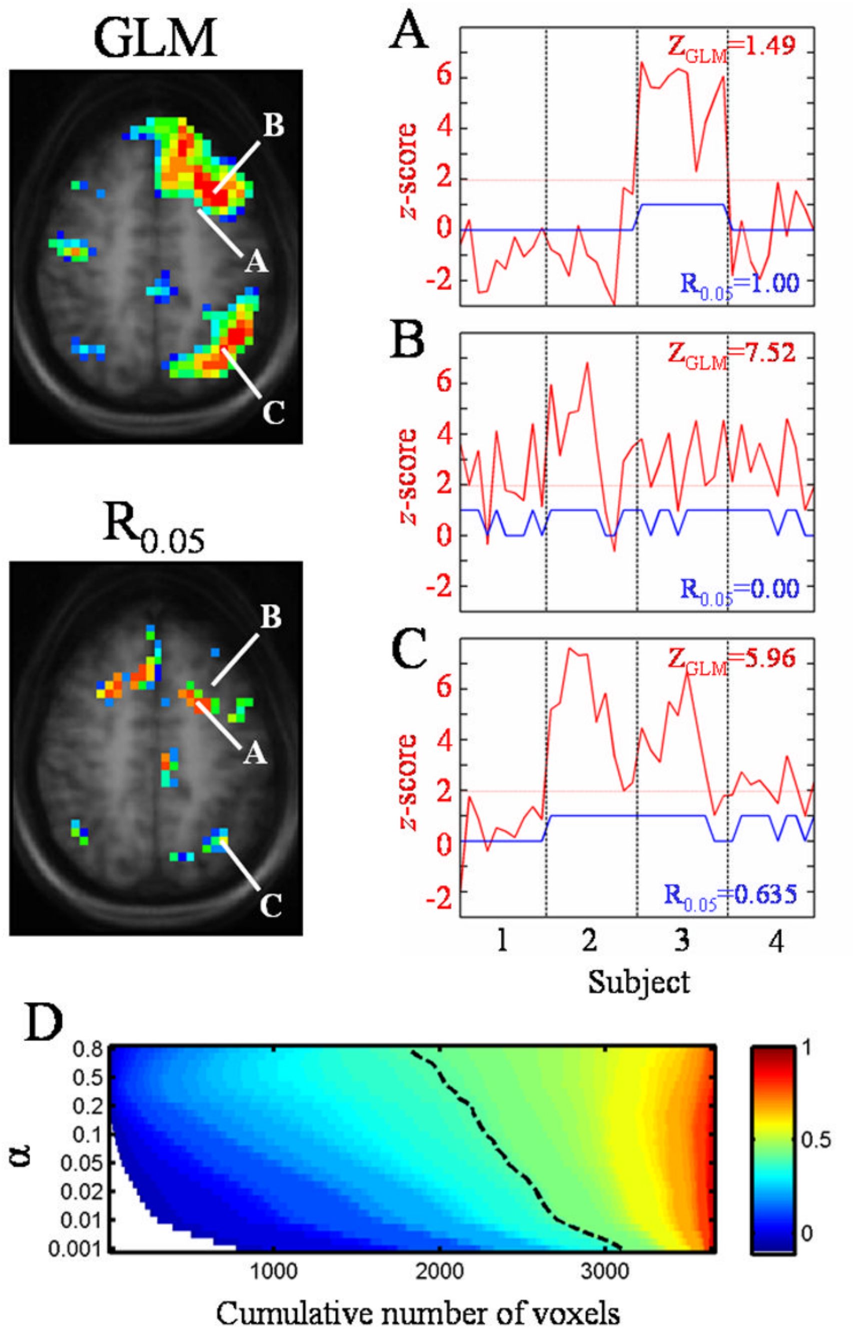

Figure 4.

Comparison of group composite and reliability maps and dependence of the ICC on α threshold. Figures on the left show slices from the group composite map (top) and reliability map (bottom) for the control subject SDTD task. Talairach coordinates of the labeled voxels are as follows: A = [−22, 7, 43], B = [−30, 15, 43], C = [−34, −61, 43]. Graphs on the right show the *z*-score across trials (red line) and the trials with significant activation (blue line, value of “0” (not significant) or “1” (significant)). The y-axis scale represents the *z*-score for the red line and is unitless for the blue line. The significance threshold of z = 1.96 (α= 0.05) is shown by the horizontal dotted line. Vertical dashed lines separate the 40 trials into groups of 10 each, one group for each subject (1 through 4, shown on bottom graph). The composite *z*-score and reliability values for each voxel appear on the graphs. The graph at bottom (D) shows the reliability (color value) as a function of α for a subset of voxels in the image. The R_α values are sorted in ascending order to show the relative number of voxels with a given reliability occurring at each α threshold. The dashed line represents $R_\alpha = 0.4$. White areas have undefined reliabilities (never significantly active).

Table 1

Subject Demographics

Group	Subject #	EHI*	Gender	Age	Lesion site	Years since stroke
Control	1	65	M	51		
	2	100	F	50		
	3	91	F	59		
	4	100	F	24		
Aphasic	1	80	M	58	Inferior frontal	8
	2	100	F	50	Temporo-parietal	7
	3	100	M	55	Inferior frontal	2
	4	100	F	50	Temporo-parietal	2

EHI — Edinburgh Handedness Inventory (for stroke subject the number reflects the pre-stroke handedness).

Table 2

Neurolinguistic and fMRI Test Performance

Group	Subject #	BDAE-3					RTT			VG[†]	SDTD[‡]
		SR	WD	WR[*]	CIM	BNT[*]	I	III[*]	VI[*]		
Control	1	—	16	5	6	15	15	15	14.6	98.8	80.2 / 99.4
	2	—	16	5	6	15	15	15	14.7	100	88.1 / 96.1
	3	—	16	5	6	15	15	15	14.9	98	64.5 / 95.8
	4	—	16	5	6	15	15	15	14.71	100	84.7 / 97.2
Aphasic	1	4[**]	16	3	6	6	15	13.88	13.83	94.8	76.5 / 87.7
	2	3	14.5	3	4	7	13.6	13.63	7.25	93.4	58.8 / 56
	3	3	15.5	4	5	10	15	13.03	11.05	98.5	74.2 / 84.5
	4	2	12	4	2	4	14.16	13.4	13.45	96.8	64.9 / 53

BDAE-3 — Boston Diagnostic Aphasia Exam-3 (Short Form); SR — Severity Rating; WD — Word Discrimination; WR — Word Repetition; CIM — Complex Ideational Material; BNT — Boston Naming Test (15 item short version); RTT — Revised Token Test (subtests I, III, and VI)

[†] VG — Verb Generation task (percentage of correct noun recalls post-scan)

[‡] SDTD — Semantic Decision/Tone Decision task (percentage of correct in-scanner responses)

[*] neurolinguistic task for which control and aphasic subjects differ significantly (P < 0.03, Mann-Whitney *U* test)

[**] dysarthria/apraxia affecting fluency/intelligibility.

Table 3

Regions of Reliable Activation in Control Subjects

Task	Brain Regions	Max. Reliability	# Voxels in ROI	ROI centroid coordinates (Talairach)
SDTD	L. GTi	0.79	18	−54, −47, −12
	R. GFi	0.67	17	32, 38, −7
	L. GTm/s	0.89[*]	13	−52, −20, 0
	R. GTm/s	0.79	26	52, −27, 5
	Th	0.67	16	−1, −6, 7
	CC, B. GC, B. PCu, Th	0.89[*]	246	−3, −37, 21
	L. Ga, L. Gsm, L. GTm/s	0.92[*]	192	−36, −54, 28
	R. GFm	0.78	29	37, 29, 22
	L. GFi/m	0.76	15	−41, 19, 28
	R. Ga, R. Gsm, R. LPi	0.73	32	36, −54, 35
	B. GFd/m/s	1.00[**]	314	−4, 21, 47
	L. Lpc	0.67	11	−5, −33, 57
VG	R. GOm, R. GTi	0.67	16	39, −59, −7
	L. GFi	0.72	71	−50, 23, 16
	R. GFi	0.90[*]	13	42, 14, 28
	L. Ga, L. GTm	0.76	10	−32, −56, 31
	L. LPi	0.89[*]	12	−41, −31, 39
	L. GFm, L. GPrC	1.00[**]	36	−44, −3, 45
	R. PCu	0.67	12	8, −38, 43
	L. GFd	0.78	31	−4, 4, 55
Combined	L. GFi	0.56	15	−54, 28, 6
	L. Ga, L. GTm	0.76	27	−29, −59, 30
	L. GFm	0.57	11	−45, 17, 28
	L. GFm	0.71	26	−36, 0, 50
	L. GFs	0.75	31	−5, 8, 54

[*] High reliability (R = 0.8—0.99)

[**] perfect reliability (R = 1.0)

L. — left; R. — right; B. — bilateral; CC — Corpus callosum; Ga — Gyrus angularis; GC — Gyrus cinguli; GFd/i/m/s — Gyrus frontalis medialis/inferior/medius/superior; GOm — Gyrus occipitalis medius; GPrC — Gyrus precentralis; Gsm — Gyrus supramarginalis; GTi/m/s — Gyrus temporalis inferior/medius/superior; Lpc — Lobulus paracentralis; LPi — Lobulus parietalis inferior; PCu — Precuneus; Th — Thalamus.

Table 4

Regions of Reliable Activation in Aphasic Subjects

Task	Brain Regions	Max. Reliability	# Voxels in ROI	ROI centroid coordinates (Talairach)
SDTD	L. Gh	0.77	31	−33, −34, −20
	R. GFm	1.00**	39	29, 39, 3
	R. GTi	0.80*	12	45, −65, −2
	L. GFi	0.78	19	−34, 37, 1
	CC, B. GC, B. PCu	0.84*	105	0, −43, 23
	L. GTs	0.80*	16	−48, −44, 19
	L. Ga, L. GTs	1.00**	48	−33, −53, 33
	L. GFi	0.65	13	−49, 3, 26
	R. GFi	0.88*	18	45, 4, 32
VG	L. GTm	0.77	10	−47, 4, −25
	R. GFi	1.00**	47	35, 22, 7
	L. GFi/m	0.68	15	−34, 38, 3
	B. GFs	0.65	10	0, 31, 49
	R. GFd	0.74	15	2, 7, 49
Combined	R. GFm	0.74	17	29, 38, 3
	L. GFm	0.73	18	−34, 38, 3

* High reliability (R = 0.8—0.99)

** perfect reliability (R = 1.0)

L. — left; R. — right; B. — bilateral; CC — Corpus callosum; Ga — Gyrus angularis; GC — Gyrus cinguli; GFd/i/m/s — Gyrus frontalis medialis/inferior/medius/superior; Gh — Gyrus parahippocampi; GTi/m/s — Gyrus temporalis inferior/medius/superior; PCu — Precuneus.