ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

NeuroImage 18 (2003) 468–482

NeuroImage

www.elsevier.com/locate/ynimg

Reproducibility of the word frequency effect: comparison of signal change and voxel counting

Michael W.L. Chee,* Hwee Ling Lee, Chun Siong Soon, Christopher Westphal, and Vinod Venkatraman

Cognitive Neuroscience Laboratory, SingHealth Research Laboratories, Singapore 169856

Received 29 April 2002; revised 15 July 2002; accepted 13 September 2002

Abstract

We determined the reproducibility of both the direction and the effect size of the word frequency effect (WFE) as it relates to associative semantic judgments. Sixteen volunteers were scanned twice. At the group level of analysis, signal change and voxel counting could both reproducibly detect the existence of a WFE. However, signal change data showed less intersession variation, particularly in the left inferior frontal gyrus. The effect size of WFE was well reproduced only with signal change measurements. In consideration of the signal change data, statistical threshold did not have a major effect on the detection or determination of the effect size. In general, while the direction of the WFE was reasonably reproducible at the individual level, the effect size was far less well reproduced. These findings suggest that with existing techniques, fMRI may be used to track changes in brain activation stemming from improvement in language proficiency at the group level but not at the individual level.
© 2003 Elsevier Science (USA). All rights reserved.

Introduction

fMRI is an attractive tool to use for the imaging of brain plasticity because of its ability to image the brain without the use of ionizing radiation. A compelling application of this capability to perform repeated measurements of neural activity (or more correctly, a surrogate of neural activity) over time is to study the changes in activation patterns that may well take place in the course of language acquisition. Such changes in brain activation patterns consequent on motor learning and acquisition of perceptual skills have been successfully documented (Karni and Bertini, 1997; Karni et al., 1995, 1998). However, prior to harnessing fMRI to evaluate language learning, it is first necessary to carefully evaluate the reproducibility of the results obtained from one scanning session to the next. (See Poldrack (2000) for an overview of the conceptual issues relating to the performance of longitudinal studies using neuroimaging.)

Previous studies examining the reproducibility of fMRI have used visual (McGonigle et al., 2000; Miki et al., 2000; Rombouts et al., 1997, 1998), motor (Cohen and DuBois, 1999; McGonigle et al., 2000; Noll et al., 1997; Ramsey et al., 1996; Tegeler et al., 1999; Yetkin et al., 1996), and various cognitive tasks involving language and working memory (Casey et al., 1998; Machielsen et al., 2000; McGonigle et al., 2000; Ojemann et al., 1998; Ojemmaan et al., 1998; Rutten et al., 2002). These studies have established that while group-level reproducibility of activation can be demonstrated, there is considerable intersession (McGonigle et al., 2000) and even intrasession (Duann et al., 2002) variability in activation.

The use of simple sensory and motor tasks could be expected to yield more consistent test-retest results than tasks tapping higher cognitive function due to relatively higher BOLD signal change and the smaller likelihood of activation in these regions being modulated by differences in processing strategy. This said, it has been shown that attention may modulate activation even in the auditory

* Corresponding author. Cognitive Neuroscience Laboratory, SingHealth Research Facilities, c/o Singapore General Hospital, 7 Hospital Drive #01-11 Singapore 169611, Singapore. Fax: +65-62246386.

E-mail address: mchee@pacific.net.sg (M.W.L. Chee).

1053-8119/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
PII: S1053-8119(02)00019-8

(Grady et al., 1997; Woodruff et al., 1996) and visual cortices (Woodruff et al., 1996).

Although spatial memory (Noll et al., 1997) and figural memory encoding (Machielsen et al., 2000) as well as some language tasks (Rutten et al., 2002) have yielded reasonably reproducible results, there is lingering concern that with tasks tapping higher cognitive abilities, volunteer skill or the strategy adopted may modulate cortical activation (Reichle et al., 2000). In particular, it is unclear if the differential response obtained when processing a particular task at two levels of difficulty can be reliably reproduced. Determining the answer to this question was the primary aim of the present investigation.

Our interest in the relative differences of cortical activation within the left prefrontal region to differential task demands is motivated by the observation that activity in this region may be modulated by semantic retrieval effort (Chee et al., 2002). More specifically, retrieval effort may be influenced by the word frequency of the test items used in a semantic judgment task when the association between test words is controlled.

Previous work using a cross-sectional design has shown that language proficiency may modulate left prefrontal activation when volunteers perform semantic associative judgments (Chee et al., 2001). Healthy volunteers who were more proficient in one language showed less prefrontal activation compared to individuals who were less proficient in that language. We posit that over time, attainment of greater proficiency in an individual's second language will result in a change in the frequency rank order of second language words in that individual's lexicon. Following from this, high frequency second language words, resembling low frequency first language words, could be expected to initially produce significantly greater activation compared to high frequency first language words. Over the course of learning, the relative difference in levels of activation between high frequency second and first language words would narrow. While an interesting proposition, the explicit demonstration of these findings in a longitudinal study is necessary before this hypothesis can be accepted. An important starting point for evaluating changes related to second language learning is to establish the reproducibility of the word frequency effect using fMRI.

A second goal of this study was to determine whether signal change or voxel counts would more consistently reproduce the word frequency effect across two scanning sessions. We previously quantified the imaging equivalent of the word frequency effect in terms of relative signal change in a functionally defined region of interest (ROI) (Chee et al., 2002). The use of signal change as a metric of activation level has been advocated as being more reliable than voxel counting (Cohen and DuBois, 1999). On the other hand, language laterality indices of brain activation derived from voxel counts have been cross-validated with the Intracarotid Amobarbital Test (Benson et al., 1999; Binder et al., 1996; Desmond et al., 1995; Lehericy et al., 2000). Furthermore, voxel counting is popular in clinical applications.

Finally, we were interested in examining how the use of different statistical thresholds would modulate group as well as individual results with signal change and voxel-counting metrics.

Methods

Behavioral task

Sixteen neurologically normal, right-handed participants (12 men and 4 women aged between 21 and 27 years) gave informed consent for this study. Participants were chosen on the basis of good performance in standardized English examinations described previously (Chee et al., 2002). After a briefing and out-of-scanner trial runs, these volunteers performed semantic associative judgments on word triplets in a block-design fMRI experiment (Fig. 1). They were instructed to choose the word from a pair that was more closely related to the sample stimulus (uppermost item in each panel) and to press the appropriate button on a two-button mouse. In this adaptation of the Pyramids and Palm Trees (PPT) task (Howard and Patterson, 1992), stimulus word triplets were designed so as to make the "correct" answer obvious. This was to reduce the confounding effect of relative association strength on retrieval effort (Fletcher et al., 2000).

Words used to create the stimulus triplets were obtained from the MRC Psycholinguistic Database (http://www.itd.clrc.ac.uk/Projects/Psych/psych.html). Two sets of word triplets were created according to methods described previously (Chee et al., 2002). Each set contained 48 high frequency and 48 low frequency triplets. High frequency words in Set A had a median frequency of 68.3 occurrences per million words (mean = 85.4, SD = 56.9) and low frequency words had a median frequency of 3.00 occurrences per million words (mean = 3.48, SD = 1.50) (Kucera and Francis, 1967). In Set B, high frequency words had a median frequency of 63.2 occurrences per million words (mean = 79.5, SD = 45.8) and low frequency words had a median frequency of 2.67 occurrences per million words (mean = 2.76, SD = 1.20). High and low frequency words were matched on concreteness, and the respective "concreteness value" means were 573 and 572 (Set A) and 572 and 569 (Set B). The semantic relatedness ratings of words in both sets of stimuli were matched in order to ensure similarity of task difficulty (Chee et al., 2002).

In the control task, the sample comprised a string of "O's" which varied in length from 3 to 6 (i.e. "OOO" to "OOOOOO"). One of a pair of "O-strings" was 6% smaller (or larger) than the sample and the other was 12% smaller (or larger) (Fig. 1). Participants were instructed to choose the item that was closer in size to the sample stimulus and to indicate their choice by pressing the right or left mouse button. This modification of the size judgment task sought

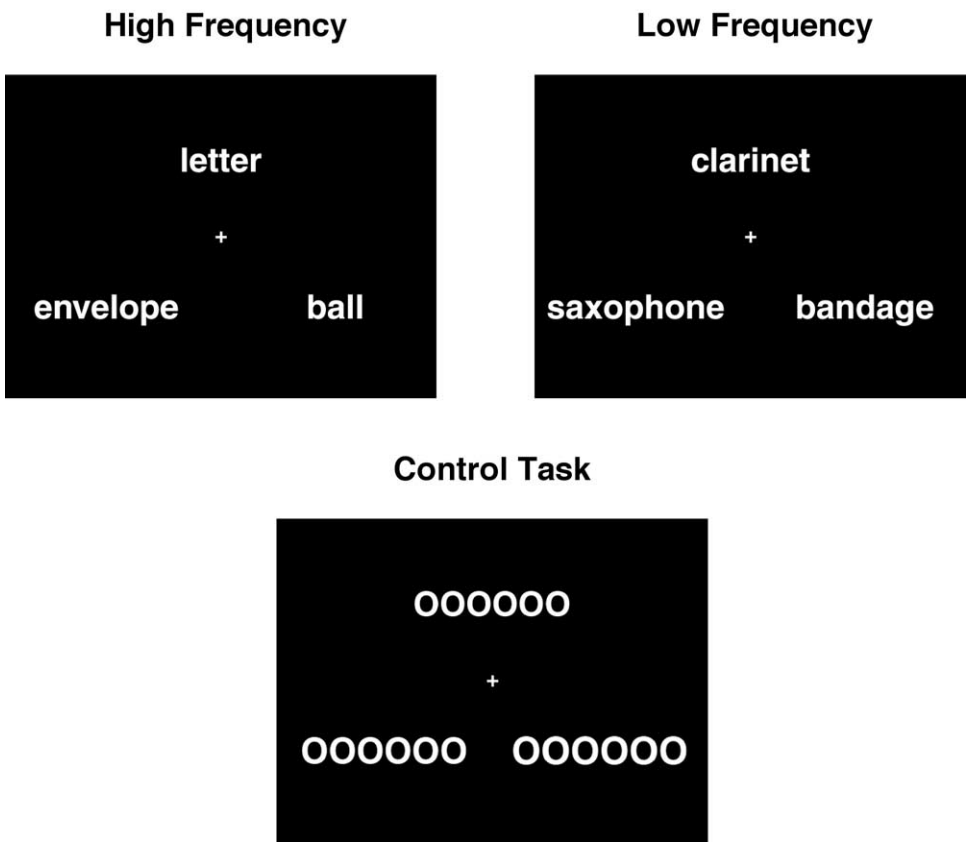


Fig. 1. Exemplars of the stimuli used in the semantic judgment (high and low frequency) and control tasks.

to remove possible confounding effects of using words in a perceptual judgment task.

High and low frequency stimuli were alternated and interleaved between control blocks. The order of presentation of blocks was counterbalanced across the two runs in this experiment. In total, there were 8 blocks of high frequency and 8 blocks of low frequency words. Each task block lasted 18 s and each control block lasted 30 s. Each stimulus appeared for 2.5 s and was followed by 0.5 s fixation. The two distinct sets of words used were counterbalanced across sessions in different volunteers.

Quality control of MR data

MR signal stability in the spatial and temporal domain were examined at the beginning of each week of the study according to a previously described quality control protocol (Weisskoff, 1996). Using a gel-based test phantom, images acquired over 400 time points showed a mean single-voxel standard deviation of the MR signal (measured at the center of the phantom) of approximately 0.64%. The F_{15} of the scanner (a measure of signal fluctuation over a 15 $\times$ 15 voxel square and expressed as a relative deviation) was between 0.15 and 0.18%.

Imaging and image analysis

Experiments were performed in a 2.0 T Bruker Tomikon S200 system (Bruker, Karlsruhe, Germany). A blipped gradient-echo EPI sequence was used with a TR of 2000 ms, a FOV of 23 $\times$ 23 cm, and a 128 $\times$ 64 pixel matrix. Fifteen oblique axial slices approximately parallel to the AC-PC line 4 mm thick (2 mm gap) were acquired. High-resolution anatomical reference images were obtained using a three-dimensional spoiled-gradient-recalled-echo sequence. A bite bar was used to reduce head motion as well as to reduce variation in head position across scanning sessions. Participants were scanned in two sessions separated by 1 week. As far as possible, the anatomical alignment used in the first session was replicated in the second session.

Following phase correction, the functional images were analyzed using Brain Voyager 2000 software version 4.6 (Brain Innovation, Maastricht, Holland). Intensity normalization was performed prior to motion correction. Gaussian filtering was applied in the temporal and spatial domains. In the spatial domain a smoothing kernel of 4 mm FWHM was used for the computation of individual activation maps while a 3 time-point Gaussian FWHM filter was used in the temporal domain. Registration of the functional MR data set to the high-resolution anatomical image of the brain was achieved by registering the functional MR data set to the stack of coplanar T2 images acquired at the end of the study and finally registering these images to the 3-D image. The resulting realigned data set was then transformed into Talairach space (Talairach and Tournoux, 1988).

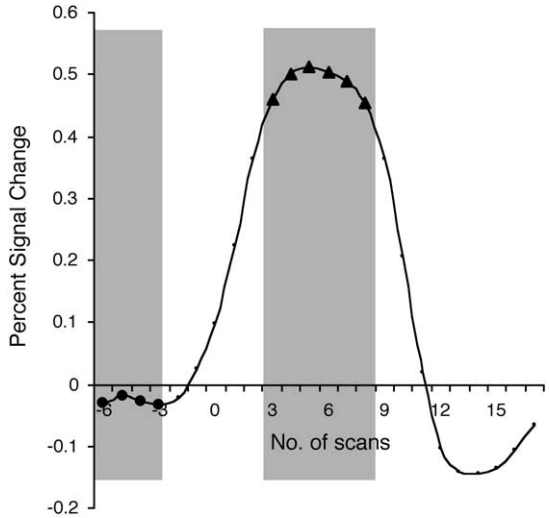


Fig. 2. An averaged block time course showing the points considered in the percentage signal change analysis. A mean percentage signal change value for each timuls type was obtained by subtracting the average signal at points from time −6 to −3 (points marked ●) from the average signal at points 3 to 8 (points marked ▲).

Individual subject statistical maps were computed using a general linear model (GLM) with two explanatory variables: high and low frequency test items. The expected BOLD signal change was modeled using a modified gamma function (Boynton et al., 1996) (tau of 2.5 s and a delta of 1.5) synchronized to blocks of cognitive tasks. Statistical maps for individual participants, from which ROI-based analysis was performed, were created using F (2,756) values greater than 8, 22, and 30, corresponding to lax, regular, and conservative statistical thresholds, respectively. The "$F > 22$" threshold used was comparable to previous studies reported by our laboratory (Chee et al., 2001, 2002).

Percentage signal change analysis: individual ROI approach

For each individual's data, regions of interest in the left prefrontal region (corresponding to Brodmann's areas 44, 45, 47, 6, and 9) encompassing the inferior and middle frontal gyri were defined by sampling volumes that were active in both low and high frequency semantic judgment relative to size judgment. The data obtained were analyzed in two "bins": left inferior frontal gyrus (LIFG) and the left middle frontal gyrus (LMFG). The motivation to partition the ROI arose from three sources. The anterior left inferior prefrontal region (corresponding to the pars triangularis and pars orbitalis portion of the LIFG) has been shown by a number of functional neuroimaging studies to play an important part in semantic processing (Poldrack et al., 1999). The LIFG has been shown to give more consistent results compared to other frontal and temporal regions in studies seeking to correlate language lateralization indices derived from fMRI and intracarotid amobarbital testing (Hund-Georgiadis et al., 2001; Lehericy et al., 2000). Finally, our own findings using the task implemented in the present study suggests that the activity of the LIFG is modulated by word frequency (Chee et al., 2002). The spatial location of peak activation was determined using Talairach Daemon (http://ric.uthscsa.edu/projects/talairachdaemon.html).

Random and unknown systematic effects could contribute to differences in spatial location of activation when high or low frequency items are individually selected as the predictor of interest. To overcome these effects, we selected ROIs jointly activated in both high and low frequency conditions as this was deemed the least biased comparison of activation between these conditions. Since a rigid anatomical template was not used (so as to account for individual variations in structural and functional anatomy) some voxels were classified under both LIFG and LMFG. This occurred in volunteers who showed extensive activation.

Within each individual's ROI, averaged time courses comprising 15 time points (9 task-related and 6 baseline points) were considered in computing the average BOLD signal change due to the semantic tasks with respect to their size judgment baseline tasks. This percentage signal change for each semantic judgment task and for each individual was calculated by subtracting the signal change corresponding to the average signal derived from the points in time − 6 to − 3 (corresponding to the size judgment task) from the average signal obtained from points 3 to 8 located on the plateau of the BOLD response curve (Fig. 2). In this way, points in the transition phase during the rise and fall of the BOLD signal were omitted.

Repeated-measures ANOVA were performed on the percentage signal change data in the LMFG and LIFG, with session (1 vs 2), word frequency (high vs low), and statistical threshold ($F > 8$ vs $F > 22$ vs $F > 30$) as within-subject variables using SPSS 10 (SPSS, Chicago, IL).

Percentage signal change analysis: group-based ROI approach

For each session, and for each of two statistical thresholds, a group-level activation map showing voxels in the left

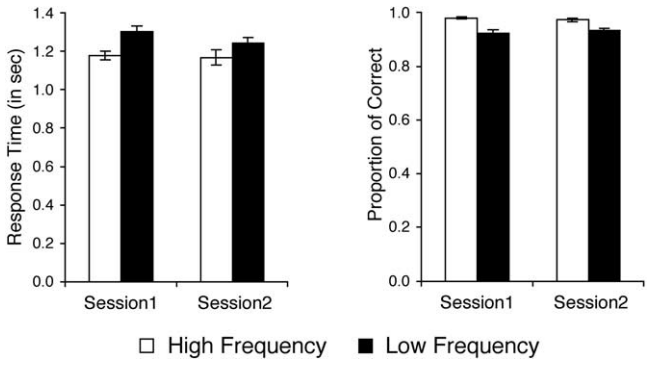
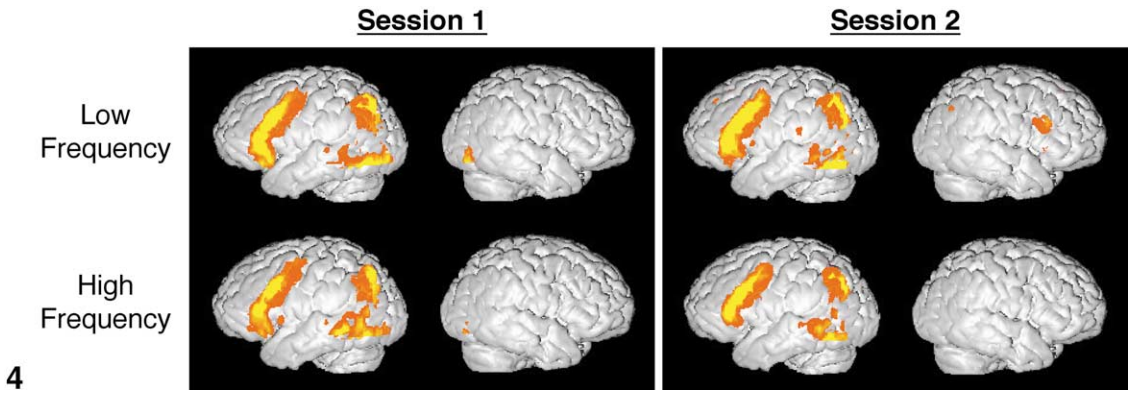


Fig. 3. Response time and accuracy data associated with semantic judgment of high and low frequency items in the two test sessions. Error bars denote 1 standard error.



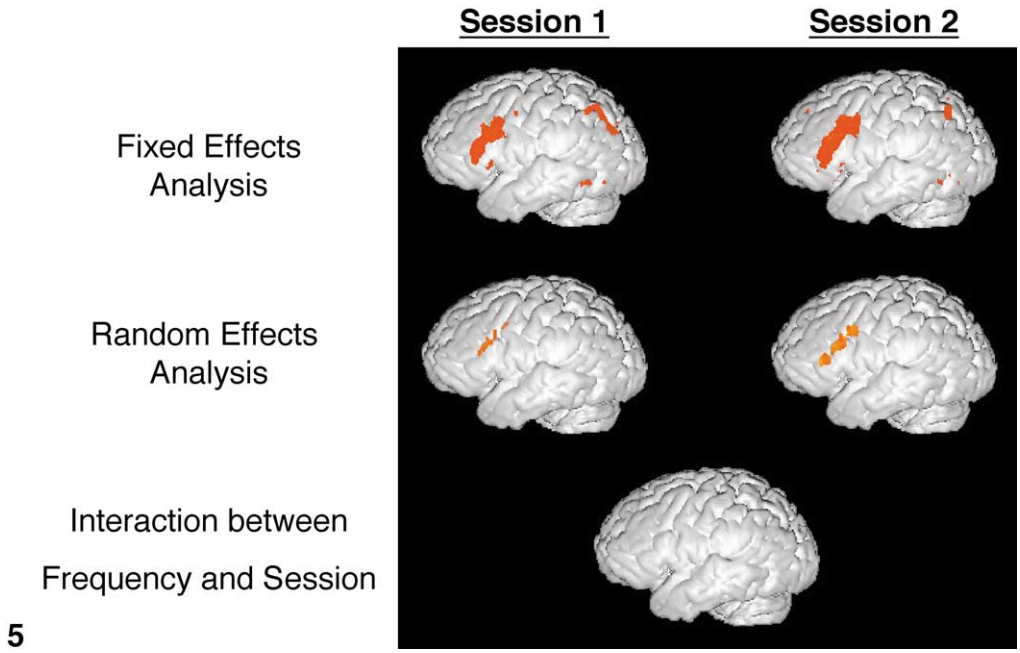


Fig. 4. Regions activated during semantic judgment involving high or low frequency items compared to size judgment in Sessions 1 and 2 (group-level data). A statistical threshold of P(corrected) < 0.001 was used in this fixed-effects analysis.

Fig. 5. Left hemisphere regions activated in the contrast between low and high frequency items (low > high) using fixed-effects [P(uncorrected) < 0.001] and random-effects analyses [P(uncorrected) < 0.001]. The lowermost figure shows the lack of interaction between frequency and session.

prefrontal region that were significantly more active in the low frequency condition compared to the high frequency condition was computed. Parameter estimates from the LIFG and LMFG were derived from the best fit of a GLM performed on these voxels. Although providing a "standardized" collection of data, this approach does not take into account deficiencies in the Talairach spatial normalization process. This method also does not account for interindividual variability in location of activation.

Voxel-count analysis

Most of the steps used in this analysis were similar to those used for percentage signal change analysis with one important difference: the ROI assessed for high frequency items and for low frequency items was that volume activated above threshold in each experimental condition. Voxel counts were performed using an automated program at each of these three different statistical thresholds.

Normalized index of difference (NID)

Normalized indices of difference (NID) of signal magnitude and voxel counts were calculated for the two item sets: $[2 * (\text{low} - \text{high})/(\text{low} + \text{high})]$, at each statistical threshold. The rationale for creating this index is as follows: Unlike PET, fMRI does not generate absolute signal magnitude values although relative signal values should be preserved. If fMRI results were truly replicable across test sessions, we would expect that the relative difference in activation associated with low frequency and high frequency words would be relatively constant. The denominator in the NID seeks to compensate for the random variation in the absolute value of the signal between test sessions.

Table 1
Talairach coordinates of activation peaks, $R_{overlap}$ and R_{size} pertaining to individual volunteers' (S1. . .S16) activation within the left middle frontal gyrus (LMFG) sorted by word frequency

No.	High frequency								Low frequency							
	Session 1			Session 2			$R_{overlap}$	R_{size}	Session 1			Session 2			$R_{overlap}$	R_{size}
	x	*y*	*z*	*x*	*y*	*z*			*x*	*y*	*z*	*x*	*y*	*z*		
S1	—	—	—	−41	5	39	0.000	0.000	−46	7	36	−46	7	39	0.417	0.998
S2	—	—	—	−43	10	33	0.000	0.000	−49	10	35	−46	10	32	0.698	0.758
S3	—	—	—	—	—	—	—	—	−52	13	36	—	—	—	0.000	0.000
S4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
S5	−49	24	27	−49	25	24	0.368	0.535	−46	25	24	—	—	—	0.000	0.000
	−31	9	36						−42	10	44					
S6	−43	33	−3	−47	10	30	0.001	0.160	−42	12	30	—	—	—	0.000	0.000
				−40	10	46			−46	20	32					
S7	−43	4	39	−39	4	39	0.604	0.772	−43	4	39	−39	4	39	0.0001	0.856
	−40	13	29	−40	16	30			−43	16	24	−41	13	30		
S8	−37	9	39	−43	10	32	0.249	0.738	−38	7	42	−46	4	45	0.409	0.888
S9	−43	16	27	−37	13	27	0.513	0.844	−46	16	27	−40	16	24	0.689	0.995
									−41	22	42	−49	16	32		
S10	−37	10	32	−37	10	32	0.525	0.632	−37	10	31	−37	10	31	0.615	0.877
S11	−46	13	35	−40	4	41	0.570	0.886	−46	15	33	−46	19	27	0.518	0.672
				−46	19	27						−40	4	39		
S12	−40	16	31	−43	28	18	0.367	0.434	−40	16	33	−43	30	16	0.266	0.384
	−46	4	48						−37	4	42					
S13	−37	19	24	−43	9	33	0.495	0.768	−40	10	32	−43	9	33	0.427	0.814
				−41	16	23										
S14	−40	10	36	−42	10	36	0.529	0.974	−40	31	20	−46	22	45	0.635	0.932
	−40	31	21	−40	31	24			−46	13	31	−44	13	33		
S15	—	—	—	—	—	—	—	—	−37	4	41	−49	6	43	0.156	0.285
									−43	16	27	−43	14	33		
									−46	30	15	−40	30	21		
S16	−28	4	39	−38	10	33	0.026	0.570	−28	4	39	−28	10	43	0.355	0.644
				−40	19	24						−40	19	24		
Mean							0.28	0.52							0.35	0.60
(SEM)							(0.07)	(0.09)							(0.07)	(0.10)

While the NID is a useful index of signal differences between experimental conditions, an important caveat should be noted. NID can take values from −2.0 to 2.0 with zero indicating the case when high and low frequency items produce an identical magnitude of activation. NID values between 0 and 2 indicate that both frequency conditions produce activation but that the magnitude of activation is higher for low frequency words. Critically, when activation does not occur or is below threshold in either condition, the value of NID becomes 2.0 or − 2.0. This can heavily bias the mean values of NID when considering group level results. To avoid this problem, we omitted from analysis individuals in whom either high frequency or low frequency words did not result in activation above detection threshold (voxel count of zero for that threshold; see Fig. 7 for a comparison of NID results where outliers were either included or excluded).

Session effects on spatial distribution of activation

At an individual level, three measures of reproducibility were determined. These were the location of peak activations, the number of activated voxels in the smaller of two overlapping activations (R_{size}), and a combination of reproducibility of the number of activated voxels and the location of activation ($R_{overlap}$). Under this framework, $R_{size} = 2*V_{smaller}/(V_{session1} + V_{session2})$, $R_{overlap} = 2*V_{overlap}/(V_{session1} + V_{session2})$. The maximum possible value that $R_{overlap}$ can reach for a given individual is R_{size} for that individual. In these metrics, $V_{overlap}$ refers to the number of overlapping voxels in Session 1 and 2, i.e., the intersection of voxels activated in the respective ROI obtained from each session. $V_{session1}$ and $V_{session2}$ refer to the number of activated voxels within the ROI in Session 1 and 2, respectively. $V_{smaller}$ was the voxel count of the smaller set of activated voxels from either Session 1 or 2. That is, if the voxel count obtained from the ROI in Session 2 was lower than the count from Session 1, $R_{size} = 2*V_{session2}/(V_{session1} + V_{session2})$. These measures have values that range from 0.0 (worst) to 1.0 (best). Researchers have previously used R_{size} and $R_{overlap}$ to measure reproducibility (Rombouts et al., 1998).

It is important to realize that these indices regarding spatial reproducibility utilize thresholded, binarized activa-

Table 2
Talairach coordinates of activation peaks, $R_{overlap}$ and R_{size} pertaining to individual volunteers' (S1. . .S16) activation within the left inferior frontal gyrus (LIFG) sorted by word frequency

No.	High frequency								Low frequency							
	Session 1			Session 2			$R_{overlap}$	R_{size}	Session 1			Session 2			$R_{overlap}$	R_{size}
	x	*y*	*z*	*x*	*y*	*z*			*x*	*y*	*z*	*x*	*y*	*z*		
S1	−37	10	24	−34	10	27	0.409	0.776	−37	10	24	−34	10	27	0.627	0.977
	−46	28	11	−46	28	13			−38	27	15	−34	27	15		
S2	−43	18	24	−41	31	12	0.410	0.551	−34	27	15	−43	37	14	0.672	0.769
	−43	31	9	−41	31	−3			−44	28	14	−40	25	0		
S3	−40	19	21	−40	16	22	0.556	0.964	−40	19	21	−39	19	24	0.454	0.664
	−37	19	−9	−40	28	15			−40	25	18	−39	28	15		
S4	−37	7	33	−37	7	30	0.667	0.853	−38	7	33	−31	28	18	0.450	0.940
	−35	25	21	−31	28	17			−34	25	21	−49	25	21		
S5	−47	28	18	−46	28	12	0.282	0.515	−53	22	8	−46	28	13	0.122	0.253
				−28	12	33						−30	10	32		
S6	−37	10	24	−37	13	24	0.708	0.895	−37	12	24	−37	13	24	0.649	0.925
	−41	28	12	−38	28	12			−41	28	14	−40	28	12		
S7	−37	25	3	−37	25	0	0.536	0.818	−46	25	15	−37	22	14	0.583	0.967
	−46	25	15						−37	25	3	−37	25	0		
S8	−40	24	9	−46	29	12	0.361	0.717	−38	25	−9	−46	28	12	0.426	0.727
	−38	25	−10	−46	16	0						−46	19	3		
S9	—	—	—	−53	22	15	0.000	0.000	−48	22	15	−55	22	15	0.430	0.767
S10	−46	33	9	−44	34	9	0.182	0.704	−52	22	18	−52	22	18	0.364	0.568
	−47	22	−6	−47	22	−6			−48	22	−6	−49	21	−6		
S11	−40	22	0	—	—	—	0.000	0.000	−41	22	0	−38	22	3	0.442	0.925
									−35	22	9					
S12	—	—	—	−49	18	22	0.000	0.000	—	—	—	−45	16	24	0.000	0.000
S13	−43	28	8	−46	25	5	0.256	0.757	−43	28	7	−46	25	3	0.503	0.953
									−43	21	2					
S14	−40	30	−3	−46	23	11	0.318	0.715	−37	24	−9	−50	19	14	0.423	0.516
	−34	22	−9	−32	25	−8						−43	31	3		
S15	−42	4	21	−40	20	3	0.077	0.858	−46	28	0	−40	21	2	0.095	0.664
	−46	28	2	−31	25	−9						−43	31	1		
S16	−39	19	21	—	—	—	0.000	0.000	−46	25	18	−49	25	22	0.259	0.328
	−49	22	21													
Mean							0.30	0.56							0.41	0.68
(SEM)							(0.06)	(0.09)							(0.05)	(0.07)

tion maps where the spread of actual signal values is ignored. Given this, it is theoretically possible for activations in two or more sessions to overlap spatially but to have separate centroids. To cater for this contingency, we tabulated the coordinates of activation peaks at the two ROIs for each volunteer and for each session.

Group level voxel-by-voxel activation maps

At the group level, activation maps were computed using fixed as well as random effects analyses. Fixed effects analysis maps showing activation associated with low and high frequency triplets relative to size judgment were computed and displayed for each session. This was intended to give the reader a clear idea of the regions that activated during the performance of the semantic association task across all test sessions. In addition, random effects analyses were run with the intention of clarifying if the inference regarding the existence of a region (or regions) sensitive to word frequency could be generalized.

Results

Behavioral results

Repeated-measures ANOVA, with session (1 vs 2) and frequency (high vs low) as within-subject variables, were performed on response times and accuracy. With respect to response times, there was only a main effect of frequency [$F(1,15) = 35.76$, $P < 0.001$], but no other main effect or interaction (Fig. 3). Although there was a main effect of word frequency on accuracy [$F(1,15) = 32.03$, $P < 0.001$], the mean performance on low frequency items was still above 85%, indicating a high level of performance.

Reproducibility of activation in the spatial domain

Visual inspection of the group level, fixed effects analysis maps for activation in response to high and low frequency words showed good overall concordance of areas activated across sessions. This was observed in the left prefrontal, left posterior middle, and inferior temporal and

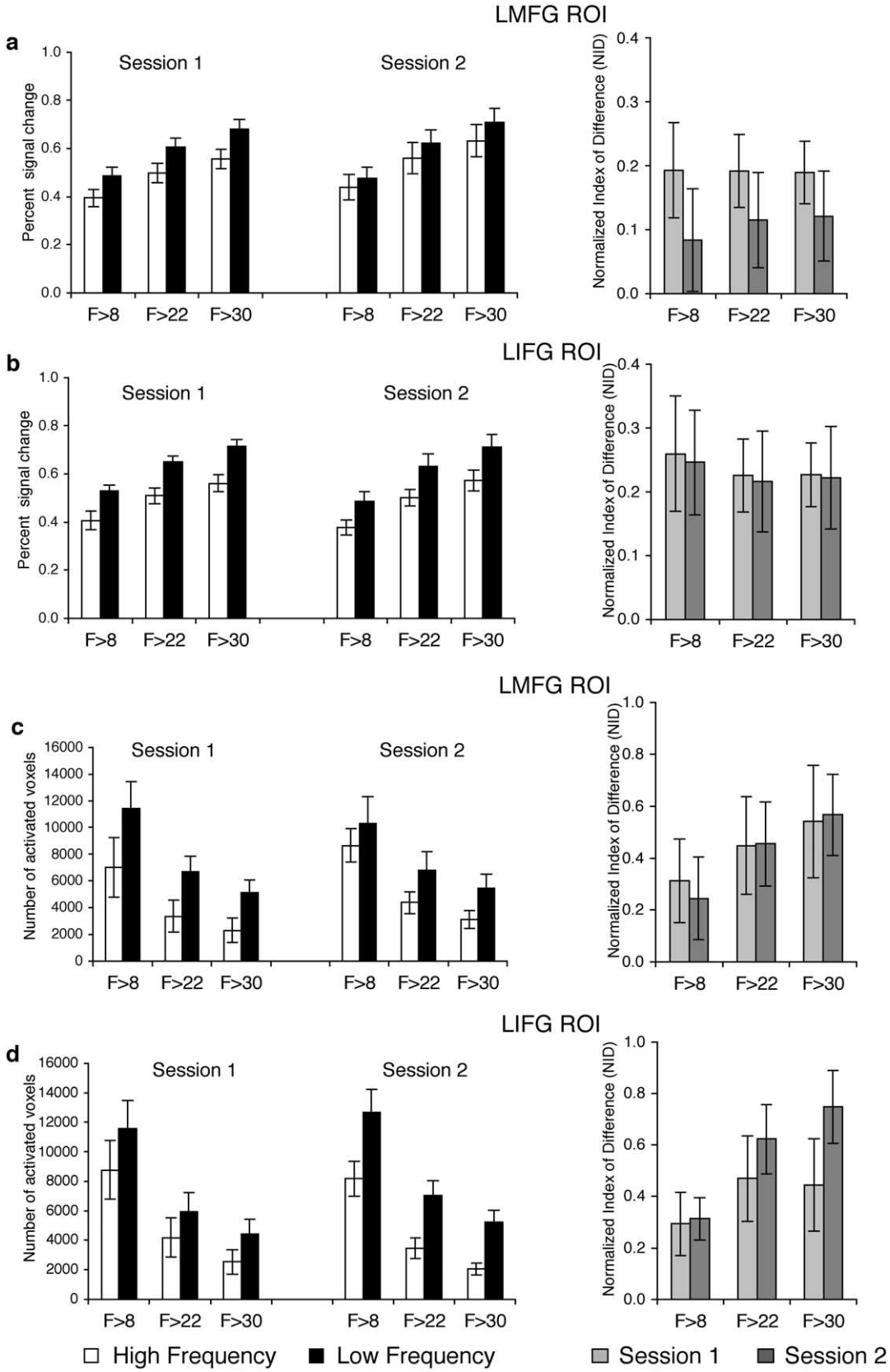


Fig. 6. Left-sided panels show the comparison of percentage BOLD signal change (a, b) and voxel counts (c, d) associated with processing high and low frequency items in Sessions 1 and 2 involving the LMFG (a, c) and LIFG (b, d). The right-sided panels show the normalized index of difference (NID) data for the corresponding test metrics, regions, and statistical thresholds. Error bars denote 1 standard error.

Table 3
Individual data showing the word frequency effect (low-high) for response time (RT), percentage signal change, and voxel-count data at threshold $F > 22$

	RT (low-high)		LMFG				LIFG			
	Session 1	Session 2	Signal change (low-high)		Voxel count (low-high)		Signal change (low-high)		Voxel count (low-high)	
			Session 1	Session 2	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
S1	0.13	0.13	0.15	0.12	2681	2253	0.23	0.19	7413	6312
S2	0.10	0.15	0.03	0.00	9209	5084	−0.05	0.28	−10821	3594
S3	0.10	0.22	0.00	0.00	4479	0	0.07	0.26	1726	10116
S4	0.10	0.08	0.31	0.29	—	—	0.26	0.34	2745	5675
S5	0.17	0.12	−0.01	0.00	376	−1004	0.00	0.06	−200	4621
S6	0.03	−0.07	0.00	0.00	4211	−2533	0.09	−0.08	1433	−682
S7	0.14	−0.11	0.10	−0.08	3392	3075	0.27	0.08	5553	3769
S8	0.21	0.12	0.20	0.13	3724	3567	0.20	0.42	2914	1723
S9	0.13	0.20	0.25	0.00	4435	5547	0.26	0.41	4630	2551
S10	0.14	0.14	0.24	0.23	7588	1516	0.09	0.16	3712	10047
S11	0.22	0.14	0.00	0.28	−389	2683	0.04	−0.14	920	812
S12	0.01	−0.02	0.03	−0.12	814	−516	0.00	−0.13	0	−2606
S13	0.08	−0.02	0.02	−0.18	29	207	0.04	−0.10	236	513
S14	0.10	0.05	0.06	−0.01	6506	8579	0.33	0.07	892	3036
S15	0.07	0.04	0.00	−0.03	2271	8184	0.00	0.05	1245	3277
S16	0.26	−0.01	0.02	0.12	434	−103	0.12	0.19	5386	4155
Mean	0.12	0.07	0.09	0.05	3317	2436	0.12	0.13	1737	3557
(SEM)	(0.02)	(0.02)	(0.03)	(0.03)	(743)	(853)	(0.03)	(0.05)	(1007)	(857)

parietal regions (Fig. 4). There was an additional area of activation above threshold in the right inferior frontal gyrus in Session 2 that was not seen in Session 1. Such additional activation has previously been observed when semantic retrieval demands are high (Fletcher et al., 2000; Roskies et al., 2001).

The random effects analysis map revealed a left prefrontal region sensitive to word frequency. The region included a superior portion of the inferior frontal gyrus and an inferior part of the middle frontal gyrus (Fig. 5). This localization was replicated across sessions and is concordant with previous findings (Chee et al., 2002).

At the group level of analysis, the measures for overlapping voxels ($R_{overlap}$) and overlapping voxels over the smaller region of activation within a specified ROI (R_{size}) were higher for the LIFG compared to the LMFG. Spatial reproducibility of activation was better for low frequency items than for high frequency items in LIFG (Table 1).

At an individual level, there was modest variability in the extent of overlap between activated voxels within the left prefrontal region across the two scanning sessions. Apart from S5, S6, S12, and S15, activation across sessions in each of the other 12 volunteers produced good to excellent spatial concordance of activation. Excluding the aforementioned four volunteers, the Euclidean distance between peak activations across condition, session, and volunteers was generally within 5–10 mm. The proportion of overlapping voxels ranged from zero (usually in the context of high frequency items and in the LMFG) to 0.7 (most commonly in the context of low frequency items in the LIFG; Tables 1 and 2).

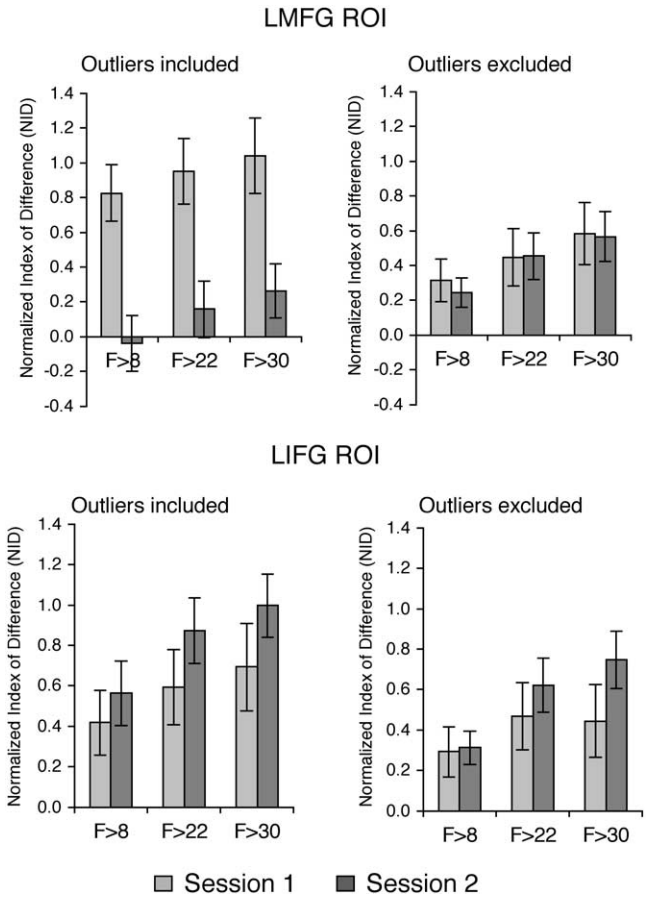


Fig. 7. Normalized index of difference (NID) for the LMFG and LIFG using the voxel counts where outliers were included (left panels) or excluded (right panels).

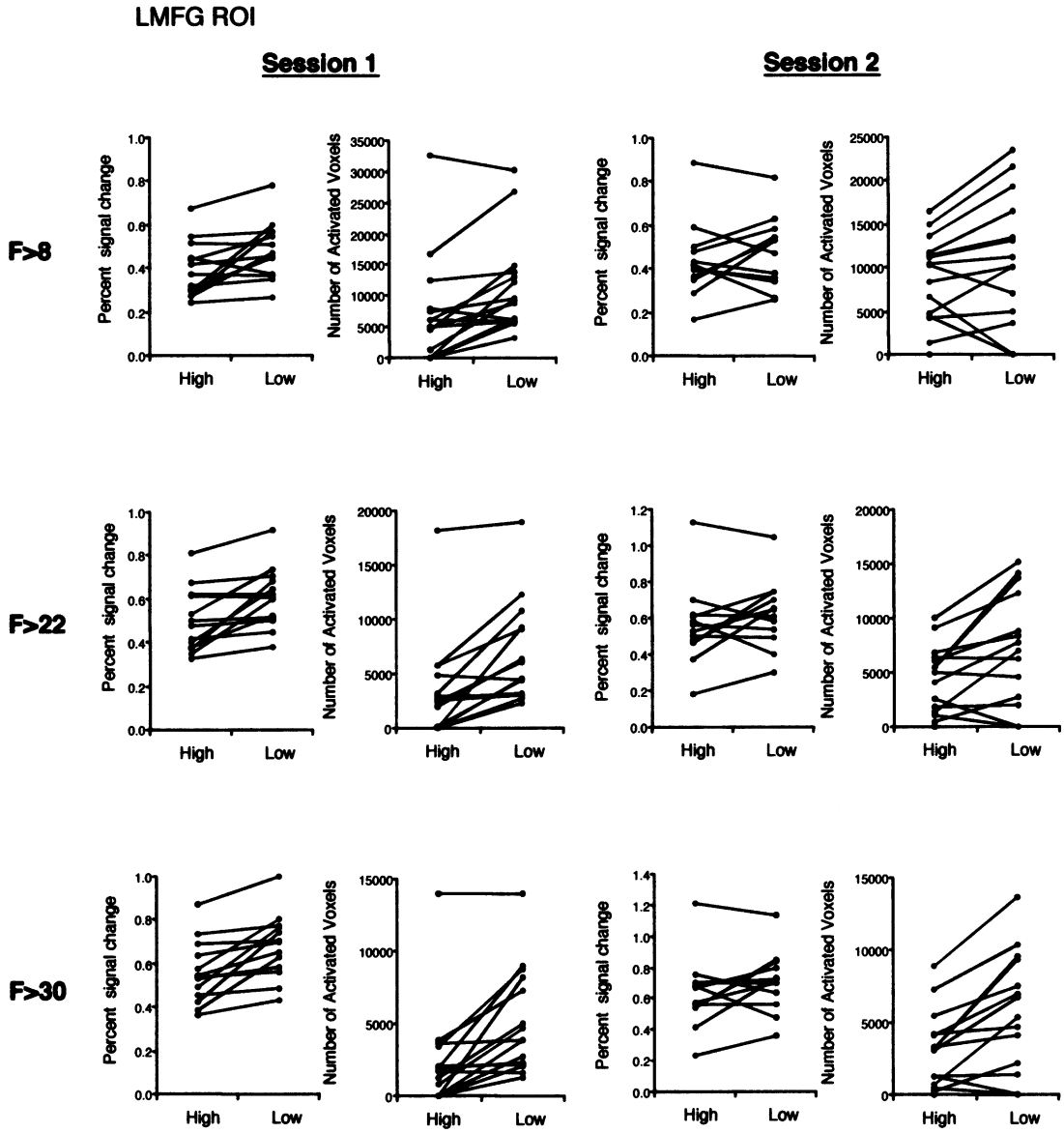


Fig. 8. Line plots showing individual BOLD signal change and voxel counts associated with the semantic processing of high and low frequency items. Analyses using different statistical thresholds in the LMFG are shown. The emphasis is on showing reproducibility of the direction of the WFE across sessions.

Reproducibility of percentage signal change analysis at different thresholds

At the group level, task-related percentage signal change in left prefrontal ROIs increased as the statistical threshold became more conservative (Fig. 6). Irrespective of the statistical threshold used, the difference in BOLD signal elicited by low and high frequency items was statistically significant across both scanning sessions in the LIFG. Within this region there were significant main effects of frequency [$F(1,15) = 17.22$, $P < 0.005$] and statistical threshold [$F(2,30) = 88.31$, $P < 0.001$] but no significant main effect of session [$F(1,15) = 1.01$, $P > 0.1$] or interaction. In the LMFG, there were significant main effects of frequency [$F(1,12) = 7.15$, $P < 0.05$] and statistical threshold [$F(2,24) = 120.88$, $P < 0.001$], but no main effect of session [$F(1,12) < 1$, n.s.]. A significant interaction was found only between frequency and statistical threshold [$F(2,24) = 10.17$, $P < 0.005$].

The difference between signal levels associated with low frequency and high frequency items reflected by NID was relatively stable across sessions and statistical thresholds for the LIFG but less so for the LMFG (Fig. 6).

At the individual level, there was substantial variation in relative signal magnitude between sessions. For example, using the "optimal" statistical threshold of $F > 22$: In Sessions 1, 12 out of 16 volunteers showed greater activation in the LIFG for low compared to high frequency items (Table 3). In 3 individuals, there was no difference in the

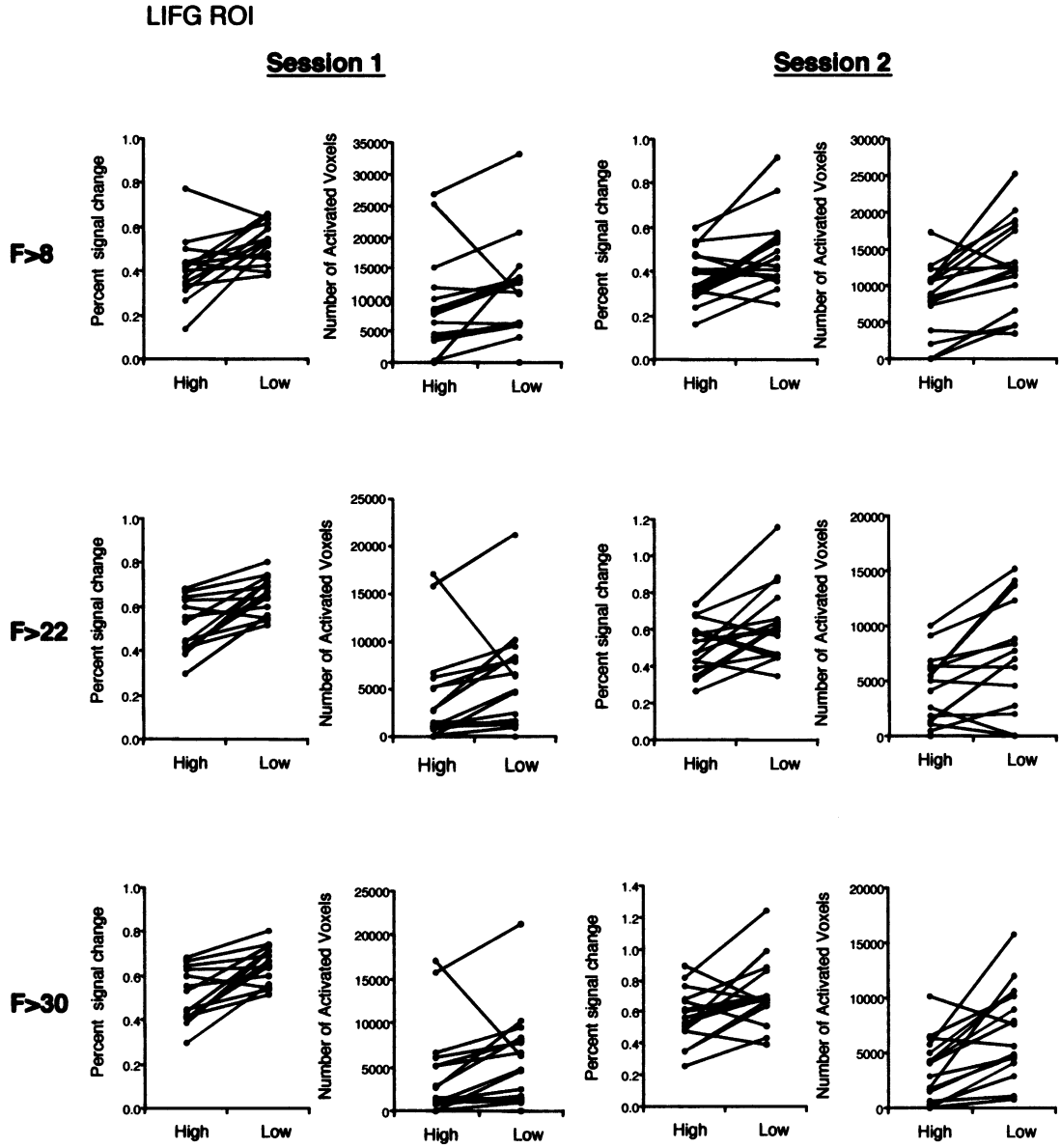


Fig. 9. Line plots showing individual BOLD signal change and voxel-count values associated with the semantic processing of high and low frequency items. Analyses using different statistical thresholds in the LIFG are shown. The emphasis is on showing reproducibility of the direction of the WFE across sessions.

magnitude of activation for low and high frequency items. One individual showed a weak effect in the opposite direction. In Session 2, 12 out of 16 volunteers showed greater activation in the LIFG for low compared to high frequency items. Four individuals showed a weak effect in the opposite direction. Nine volunteers showed changes in the same direction across the two sessions. A more complete representation of these data is shown in Figs. 8 to 11.

Statistical threshold had a modest effect on individual results as regards direction of effect and effect size. This was particularly so for data relating to the LIFG. While this conclusion can also be drawn from noting the relative stability of NID across different thresholds, it is much more impressive when inspecting the individual line graphs denoting relative activation during evaluation of low and high frequency items (Figs 8–11). Specifically, apart from one individual at the $F > 8$ threshold, the other 15 individuals data showed the same trend across the three statistical thresholds.

Reproducibility of voxel-count data

At the group level, voxel counts in left prefrontal ROIs were lower with more conservative statistical thresholds (Figs. 6c,d). Within the LMFG, there was a pronounced word frequency effect at all statistical thresholds in Session 1 and at the two higher thresholds in Session 2 (Fig. 6c). There were main effects of frequency [$F(1,14) = 16.71$, $P < 0.005$] and statistical threshold [$F(2,28) = 30.57$ $P < 0.001$] but there was no main effect of session [$F(1,14) < 1$, n.s.]. There was a significant interaction among session,

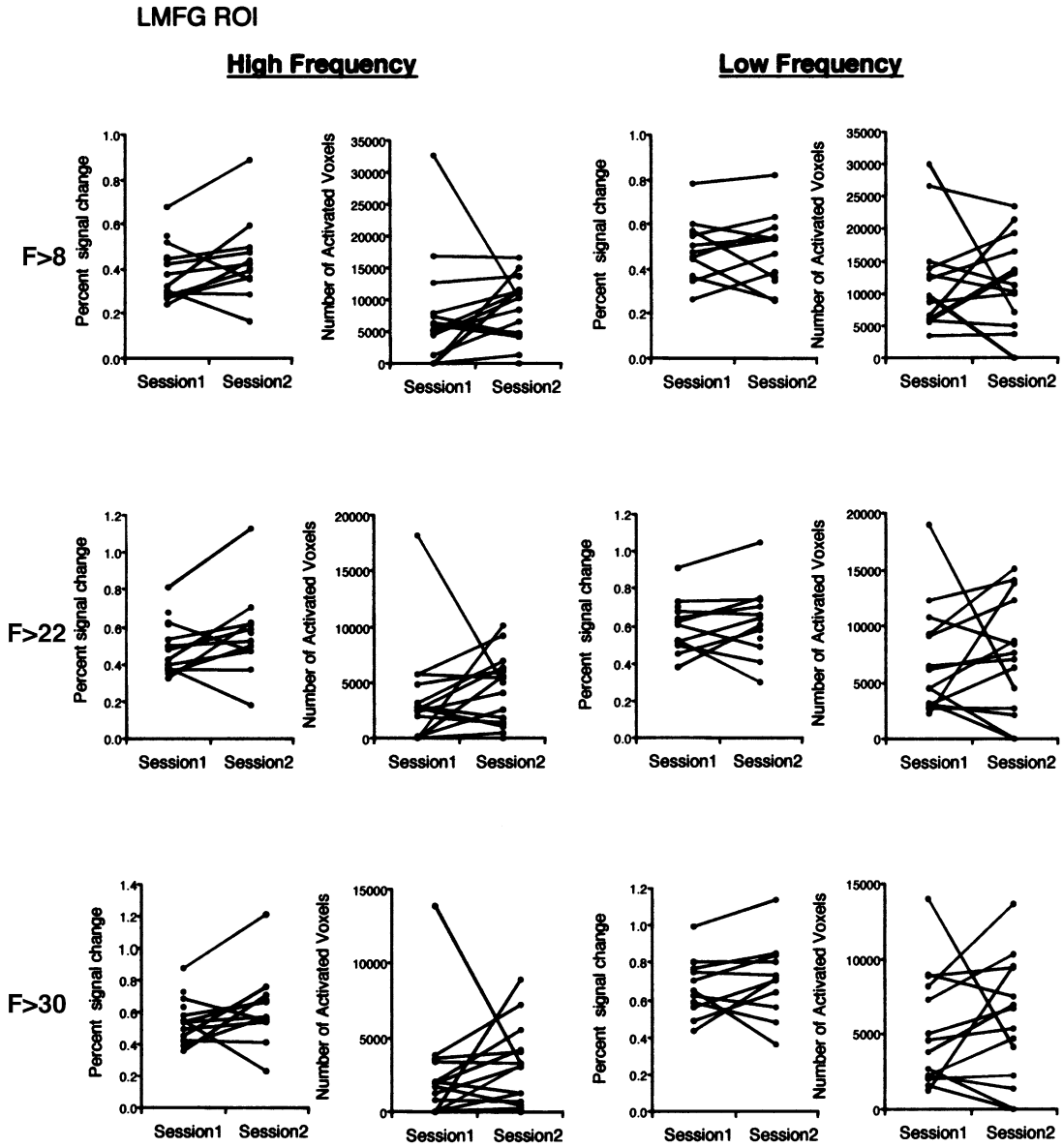


Fig. 10. Line plots showing individual BOLD signal change and voxel-count values associated with the semantic processing of high and low frequency items. Analyses using different statistical thresholds in the LMFG are shown. The emphasis is on showing reproducibility of the absolute magnitude of activation for each condition in Session 1 and Session 2.

frequency, and threshold [$F(2,28) = 5.00$, $P < 0.05$] and no other interaction. In comparison to the percentage signal change data, the interindividual variation in voxel counts was considerably higher. The difference between voxel counts associated with low frequency and high frequency items reflected by NID was more variable than the results obtained from the signal change data. The presence of outlier values in the voxel-counting data had a profound effect in reducing the reproducibility of NID. This is illustrated in the comparison in NID values where outlier data were either included or omitted (Fig. 7).

Within the LIFG (Fig. 6d), a statistically significant difference between voxel counts associated with the low and high frequency items was consistently reproduced only at the $F > 30$ threshold in Session 1, although it appeared at all statistical thresholds in Session 2. In detail, there were significant main effects of frequency [$F(1,15) = 18.45$, $P < 0.005$] and statistical threshold [$F(2,30) = 122.07$, $P < 0.001$], but no main effect of session [$F(1,15) < 1$, n.s.]. A significant interaction was observed only between frequency and statistical threshold [$F(2,30) = 3.63$, $P < 0.05$].

At an individual level, the fluctuations in voxel counts across sessions showed both volunteer and session variability in both LIFG and LMFG. This variation was more pronounced than that seen using signal change as a metric. Direction as well as effect size varied considerably between

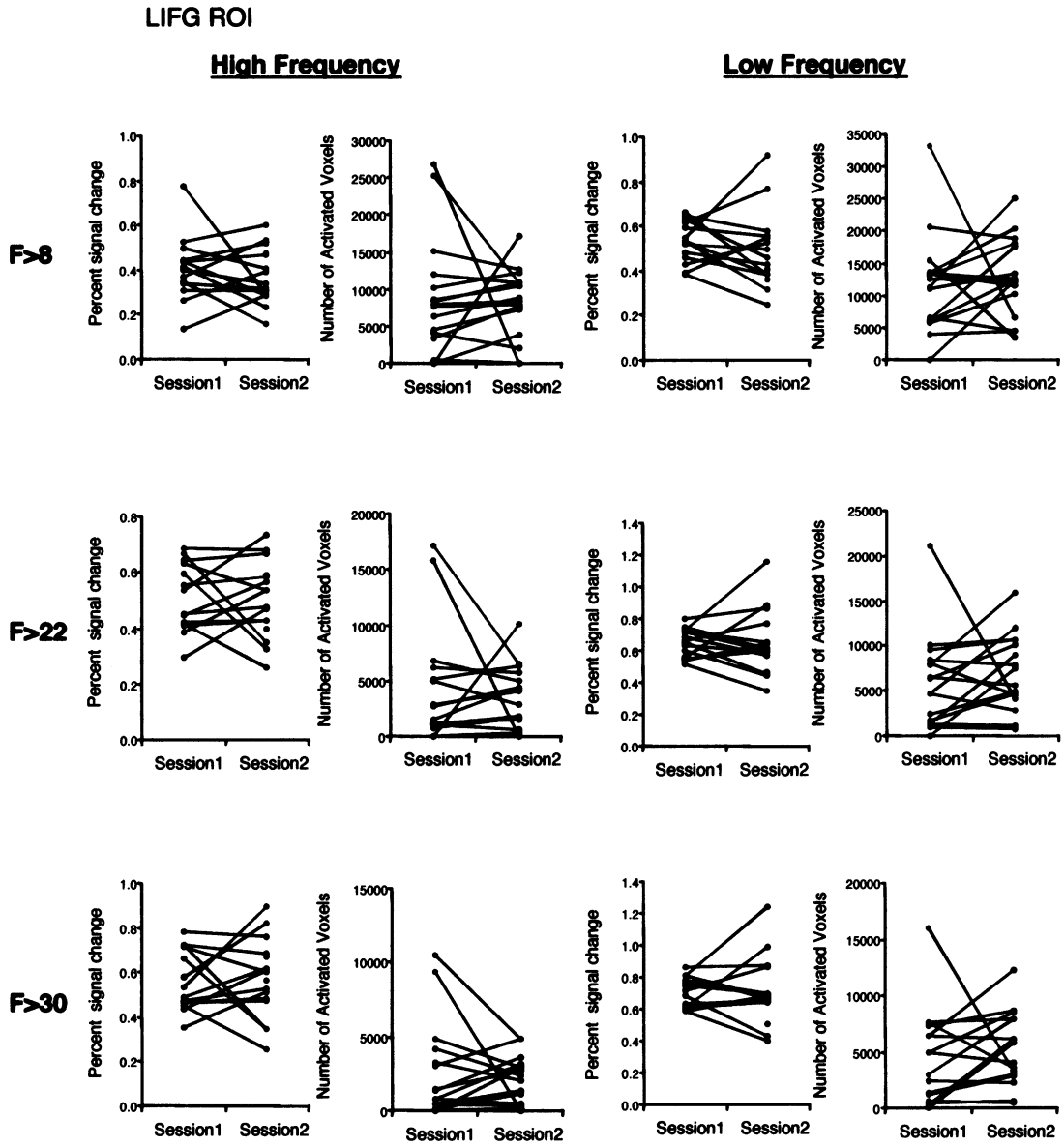


Fig. 11. Line plots showing individual BOLD signal change and voxel-count values associated with the semantic processing of high and low frequency items. Analyses using different statistical thresholds in the LIFG are shown. The emphasis is on showing reproducibility of the absolute magnitude of activation for each condition in Session 1 and Session 2.

sessions and with different thresholds as can be seen in Figs. 8–11.

Signal change using group-based and individually derived ROI

A word frequency effect was evident in both LIFG and MIFG and across sessions whichever method or statistical threshold was used. However, the reproducibility of effect size estimated from the group-based ROI was poorer compared to data obtained from individually defined ROI (Table 4). This was true of the analysis based on lower as well as higher statistical thresholds. This was also true for both prefrontal regions: the LIFG and the LMFG.

Discussion

We studied the reproducibility of fMRI data generated in an experiment that engaged higher cognitive processing. The present study adds to the considerable literature concerning this topic in several ways. The availability of concurrent behavioral information is a significant feature because it provides assurance that volunteers were engaged in the desired activity to a comparable extent across sessions. We concurrently reviewed the reproducibility of spatial and effect size data. Verifying reasonable spatial concordance of activation is an important prelude to comparing magnitude of effect since it is meaningless to compare signal magni-

Table 4
Signal change data at different statistical thresholds providing a comparison of group-based ROI (A) and individually derived ROI (B) approaches

A	LMFG				LIFG			
	$t(6110) > 3.2$		$t(6110) > 4.0$		$t(6110) > 3.2$		$t(6110) > 4.0$	
	Low	High	Low	High	Low	High	Low	High
Session 1	0.46	0.31	0.53	0.36	0.36	0.25	0.36	0.29
Session 2	0.45	0.38	0.48	0.40	0.43	0.33	0.52	0.41
B	LMFG				LIFG			
	$F(2,756) > 8$		$F(2,756) > 30$		$F(2,756) > 8$		$F(2,756) > 30$	
	Low	High	Low	High	Low	High	Low	High
Session 1	0.45	0.37	0.63	0.52	0.50	0.38	0.67	0.52
(SEM)	(0.05)	(0.04)	(0.06)	(0.06)	(0.04)	(0.05)	(0.05)	(0.05)
Session 2	0.41	0.38	0.61	0.54	0.49	0.38	0.71	0.57
(SEM)	(0.06)	(0.06)	(0.08)	(0.08)	(0.04)	(0.03)	(0.05)	(0.04)

tudes of spatially disparate networks of neurons. We documented technical parameters (Weisskoff, 1996) that could possibly affect data reproducibility as poor signal stability may contribute to Type II errors by increasing signal variance within and between sessions. We analyzed data at an individual as well as at a group level using the two most widely used metrics to quantify activation: signal change and voxel counting.

We found that while both signal change and voxel counts revealed the WFE, signal change was associated with less variability in intersession activation results and was also more robust at different statistical thresholds. As in a previous study, we found that the variation in voxel counts was up to an order of magnitude greater than the variation in percentage signal change (Cohen and DuBois, 1999).

With an ideal test metric, a "normalized" index of activation difference such as the NID would be stable across both sessions and statistical thresholds. This ideal was most closely realized with percentage signal change measurements involving the LIFG. Minimization of intersession variation of a test metric is important if the goal is to track changes in relative differences in activation, especially when the ultimate objective is to use differential responses to frequency-matched words in the native and second languages as a means of tracking learning-related neural reorganization.

The results of the present experiment further demonstrate that despite concerns about possible between-session differences resulting from differences in strategy employed and in attention, it is possible to achieve replicable results of a task involving higher cognitive processing at a group level. This stated, the variability of fMRI signal within individuals across sessions is such that studies seeking to track longitudinal changes in activation within individual are probably not feasible. Within-subject, between-session variability in results occurred despite paying careful attention to baseline signal stability, motion reduction, using a similar head orientation across sessions, and demonstrating reasonable replication of the behavioral data.

A recent study (McGonigle et al., 2000) suggested that even five retest sessions might not be sufficient to determine activation profiles that are consistently reproducible and generalizable. The present data give reason to be more optimistic for group-level data but the cautionary note struck previously should be heeded given the magnitude of within-subject, between-session variability. In addition, word lists required for longitudinal studies of language must be carefully constructed on the basis of linguistic constraints (e.g., frequency, concreteness, letter length, relatedness). As such, the number of scanning sessions that can be run is necessarily limited. Researchers will therefore have to carefully consider how to balance the number of scanning sessions run with how confidently they wish to make inferences concerning learning-related effects.

In summary, the present study represents a realistic proof-of-concept as regards the use of WFE as a means to track changes in brain topography arising from second language learning when percentage signal change in the LIFG is used as the test metric and when group level data involving a sufficient number of volunteers is available.

Acknowledgments

This work was supported by NMRC Grant 2000/0477 and BMRC Grant 257112 and 014.

References

Benson, R.R., FitzGerald, D.B., LeSueur, L.L., Kennedy, D.N., Kwong, K.K., Buchbinder, B.R., Davis, T.L., Weisskoff, R.M., Talavage, T.M., Logan, W.J., et al., 1999. Language dominance determined by whole brain functional MRI in patients with brain lesions. Neurology 52, 798–809.

Binder, J.R., Swanson, S.J., Hammeke, T.A., Morris, G.L., Mueller, W.M., Fischer, M., Benbadis, S., Frost, J.A., Rao, S.M., Haughton, V.M., 1996. Determination of language dominance using functional MRI: a comparison with the WADA test. Neurology 46, 978–984.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. 16, 4207–4221.

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., et al., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. Neuroimage 8, 249–261.

Chee, M.W., Hon, N., Lee, H.L., Soon, C.S., 2001. Relative language proficiency modulates BOLD signal change when bilinguals perform semantic judgments. Neuroimage 13, 1155–1163.

Chee, M.W.L., Hon, N.H.H., Caplan, D., Lee, H.L., Goh, J., 2002. Frequency of concrete words modulates prefrontal activation during semantic judgments. Neuroimage 16, 259–268.

Cohen, M.S., DuBois, R.M., 1999. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J. Magn. Reson Imaging. 10, 33–40.

Desmond, J.E., Sum, J.M., Wagner, A.D., Demb, J.B., Shear, P.K., Glover, G.H., Gabrieli, J.D., Morrell, M.J., 1995. Functional MRI measurement of language lateralization in Wada-tested patients. Brain 118, 1411–1419.

Duann, J.R., Jung, T.P., Kuo, W.J., Yeh, T.C., Makeig, S., Hsieh, J.C., Sejnowski, T.J., 2002. Single-trial variability in event-related BOLD signals. Neuroimage 15, 823–835.

Fletcher, P.C., Shallice, T., Dolan, R.J., 2000. "Sculpting the response space"—An account of left prefrontal activation at encoding. Neuroimage 12, 404–417.

Grady, C.L., Van Meter, J.W., Maisog, J.M., Pietrini, P., Krasuski, J., Rauschecker, J.P., 1997. Attention-related modulation of activity in primary and secondary auditory cortex. NeuroReport 8, 2511–2516.

Howard, D., Patterson, K., 1992. The Pyramid and Palm Trees Test: a test of semantic access from words and pictures. Thames Valley Test Company, Bury St. Edmunds.

Hund-Georgiadis, M., Lex, U., von Cramon, D.Y., 2001. Language dominance assessment by means of fMRI: contributions from task design, performance, and stimulus modality. J. Magn. Reson. Imaging 13, 668–675.

Karni, A., Bertini, G., 1997. Learning perceptual skills: behavioral probes into adult cortical plasticity. Curr. Opin. Neurobiol. 7, 530–535.

Karni, A., Meyer, G., Jezzard, P., Adams, M.M., Turner, R., Ungerleider, L.G., 1995. Functional MRI evidence for adult motor cortex plasticity during motor skill learning. Nature 377, 155–158.

Karni, A., Meyer, G., Rey-Hipolito, C., Jezzard, P., Adams, M.M., Turner, R., Ungerleider, L.G., 1998. The acquisition of skilled motor performance: fast and slow experience-driven changes in primary motor cortex. Proc. Natl. Acad. Sci. USA 95, 861–868.

Kucera, H., Francis, W.N., 1967. Computational Analysis of Present Day American English. Brown Univ. Press, Providence, RI.

Lehericy, S., Cohen, L., Bazin, B., Samson, S., Giacomini, E., Rougetet, R., Hertz-Pannier, L., Le Bihan, D., Marsault, C., Baulac, M., 2000. Functional MR evaluation of temporal and frontal language dominance compared with the Wada test. Neurology 54, 1625–1633.

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.

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 intersession differences. Neuroimage 11, 708–734.

Miki, A., Raz, J., van Erp, T.G., Liu, C.S., Haselgrove, J.C., Liu, G.T., 2000. Reproducibility of visual activation in functional MR imaging and effects of postprocessing. Am. J. Neuroradiol. 21, 910–915.

Noll, D.C., Genovese, C.R., Nystrom, L.E., Vazquez, A.L., Forman, S.D., Eddy, W.F., Cohen, J.D., 1997. Estimating test-retest reliability in functional MR imaging. II: application to motor and cognitive activation studies. Magn. Reson. Med. 38, 508–517.

Ojemann, J.G., Buckner, R.L., Akbudak, E., Snyder, A.Z., Ollinger, J.M., McKinstry, R.C., Rosen, B.R., Petersen, S.E., Raichle, M.E., Conturo, T.E., 1998. Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. Hum. Brain Mapp. 7, 234–243.

Ojemmann, J.G., Buckner, R.L., Akbudak, E., Snyder, A.Z., Ollinger, J.M., McKinstry, R.C., Rosen, B.R., Petersen, S.E., Raichle, M.E., Conturo, T.E., 1998. Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. Hum. Brain Mapp. 7, 234–243.

Poldrack, R.A., 2000. Imaging brain plasticity: Conceptual and methodological issues—A theoretical review. Neuroimage 12, 1–13.

Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. Neuroimage 10, 15–35.

Ramsey, N.F., Tallent, K., Vangelderen, P., Frank, J.A., Moonen, C.T.W., Weinberger, D.R., 1996. Reproducibility of the human 3D fMRI brain maps acquired during a motor task. Hum. Brain Mapp. 4, 113–121.

Reichle, E.D., Carpenter, P.A., Just, M.A., 2000. The neural bases of strategy and skill in sentence-picture verification. Cogn. Psychol. 40, 261–295.

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.

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

Roskies, A.L., Fiez, J.A., Balota, D.A., Raichle, M.E., Petersen, S.E., 2001. Task-dependent modulation of regions in the left inferior frontal cortex during semantic processing. J. Cogn. Neurosci. 13, 829–843.

Rutten, G.J.M., Ramsey, N.F., Rijen, P.C.v., Veelen, C.W.M.v., 2002. Reproducibility of fMRI-determined language lateralization in individual subjects. Brain Lang 80, 421–437.

Talairach, J., Tournoux, P., 1988. Co-planar stereotaxic atlas of the human brain. Thieme, New York.

Tegeler, C., Strother, S.C., Anderson, J.R., Kim, S.G., 1999. Reproducibility of BOLD-based functional MRI obtained at 4 T. Hum. Brain Mapp. 7, 267–283.

Weisskoff, R.M., 1996. Simple measurement of scanner stability for functional NMR imaging of activation in the brain. Magn. Reson. Imaging 36, 643–645.

Woodruff, P.W., Benson, R.R., Bandettini, P.A., Kwong, K.K., Howard, R.J., Talavage, T., Belliveau, J., Rosen, B.R., 1996. Modulation of auditory and visual cortex by selective attention is modality-dependent. NeuroReport 7, 1909–1913.

Yetkin, F.Z., McAuliffe, T.L., Cox, R., Haughton, V.M., 1996. Test-retest precision of functional MR in sensory and motor task activation. AJNR Am. J. Neuroradiol. 17, 95–98.