NeuroImage **11,** 708–734 (2000)
doi:10.1006/nimg.2000.0562, available online at http://www.idealibrary.com on IDEAL®

Variability in fMRI: An Examination of Intersession Differences

D. J. McGonigle, A. M. Howseman, B. S. Athwal, K. J. Friston, R. S. J. Frackowiak, and A. P. Holmes[1]

Wellcome Department of Cognitive Neurology, Institute of Neurology, London WC1N 3BG, United Kingdom

Received September 22, 1999

The results from a single functional magnetic resonance imaging session are typically reported as indicative of the subject's functional neuroanatomy. Underlying this interpretation is the implicit assumption that there are no responses specific to that particular session, i.e., that the potential variability of response between sessions is negligible. The present study sought to examine this assumption empirically. A total of 99 sessions, comprising 33 repeats of simple motor, visual, and cognitive paradigms, were collected over a period of 2 months on a single male subject. For each paradigm, the inclusion of session-by-condition interactions explained a significant amount of error variance ($P < 0.05$ corrected for multiple comparisons) over a model assuming a common activation magnitude across all sessions. However, many of those voxels displaying significant session-by-condition interactions were not seen in a multisession fixed-effects analysis of the same data set; i.e., they were not activated on average across all sessions. Most voxels that were both significantly variable and activated on average across all sessions did not survive a random-effects analysis (modeling between-session variance). We interpret our results as demonstrating that correct inference about subject responses to activation tasks can be derived through the use of a statistical model which accounts for both within- and between-session variance, combined with an appropriately large session sample size. If researchers have access to only a single session from a single subject, erroneous conclusions are a possibility, in that responses specific to this single session may be claimed to be typical responses for this subject. © 2000 Academic Press

INTRODUCTION

This study assesses the generality of results obtained from a single functional magnetic resonance imaging (fMRI) session. fMRI is a noninvasive technique that has revolutionized the study of human brain function (e.g., Belliveau *et al.*, 1991; Ogawa *et al.*, 1992; Kwong *et al.*, 1992). As with many brain imaging techniques, such as positron emission tomography, a number of observations (*scans*) from each subject are collected. A single experimental examination of one subject in this fashion constitutes a *session*. Although exceptions exist, it is unusual for a subject to be scanned on more than one occasion and, more often than not, a single fMRI session is assumed to give an accurate representation of a subject's functional neuroanatomy.

There may be problems, however, if one adopts this "one subject, one session" approach to neuroimaging experiments. One session is only a single, discrete "snapshot" of the subject's brain and may not epitomize responses to the sensorimotor or cognitive challenge employed. Indeed, differences between sessions are inevitable: for example, the BOLD response is an indirect and semiqualitative measure of neuronal activity, and the relationship between BOLD contrast and cerebral oxygen metabolism is influenced by a number of physiological factors (e.g., for review see Ogawa *et al.*, 1998). Furthermore, single-session results may be influenced by slight variations in the hardware characteristics of the MR scanner, which are not systematic across sessions (e.g., the shim performed to homogenize the B_0 field of the scanner; Howseman *et al.*, 1998). Any differences in subject position within the headcoil on separate scanning sessions may also result in greater variability in voxel signal changes, due to partial volume effects, as may different patterns of subject movement between sessions. In addition to the above, nonspecific physiological effects such as the level of arousal may further influence the neurovascular response to the activation task in question.

These effects are hard to control and may substantially influence a single session's results, such that the experiment may ultimately say as much about the context under which the data were acquired as the effects of the experimental manipulation itself. Although few researchers would expect a *precise* replication of the results if an experiment were repeated, it is currently unclear how generalizable single-session results are with fMRI.

[1] Current address: Robertson Centre for Biostatistics, Boyd Orr Building, University Avenue, Glasgow G12 8QQ, Scotland, UK.

1053-8119/00 $35.00

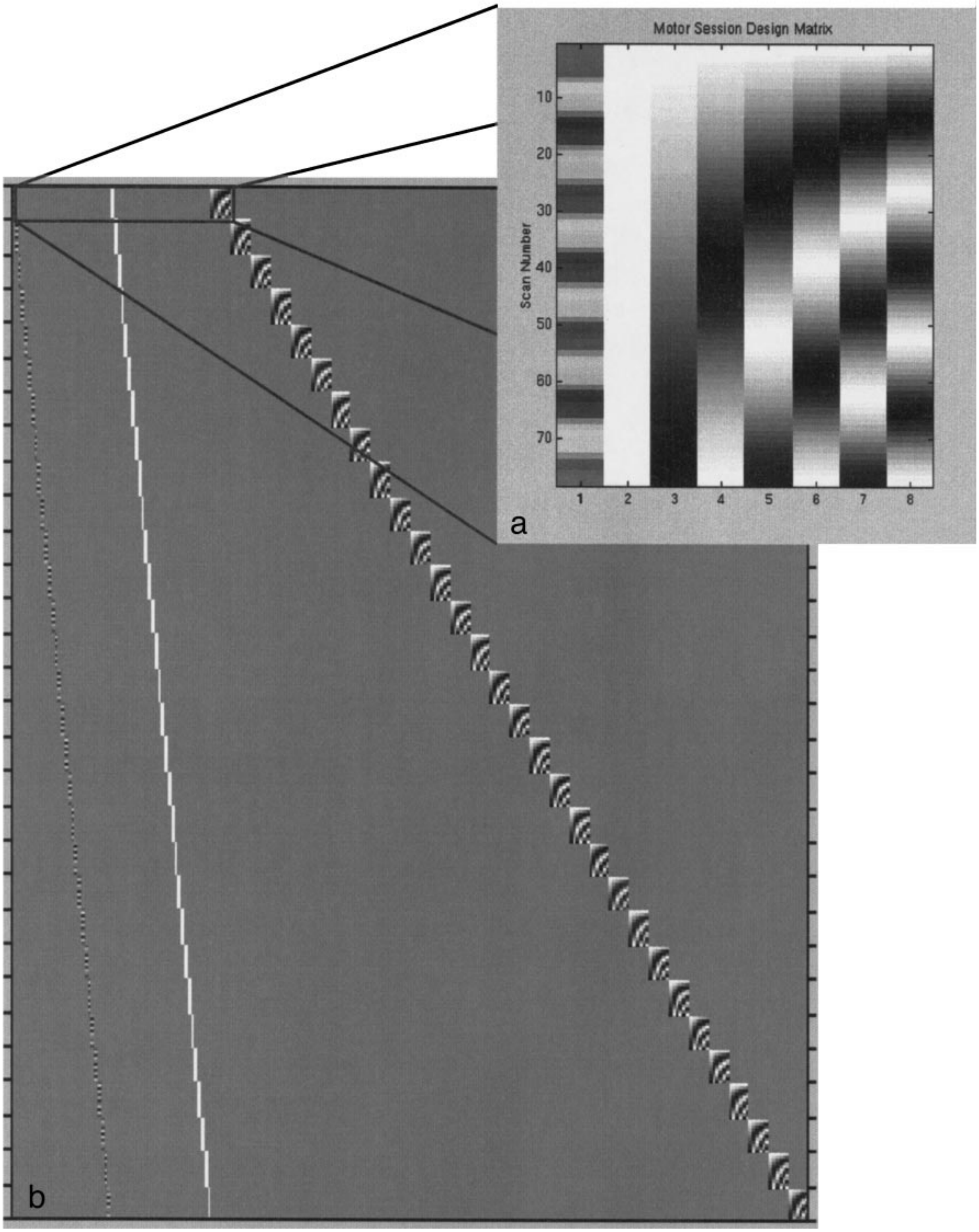

FIG. 1. Design matrices used for analysis. The design matrix is a graphical representation of the experimental model. Each column of the design matrix represents a separate regressor within the statistical model, and each row represents a single fMRI volume. The gray-scale color value within each cell displays the value of the relevant regressor at that point in the fMRI time series before the model fit is estimated. For example, a simple boxcar regressor, before convolution, would occupy a single design matrix column, and each cell within the column would have a value of 0 or 1, depending on the experimental design used. (A) A single session design matrix with the regressor of interest (the CBC, column 1), the session mean effect (γ_i, column 2), and the set of discrete cosine basis functions used to effect high-pass filtering (columns 3–8). (B) A multisession design matrix, constructed from *n* single-session design matrices, where *n* is the number of sessions analyzed at the multisession level. The design matrix in A was used only for single-session analyses, whereas B was used for both the fixed- and the random-effects multisession analyses.

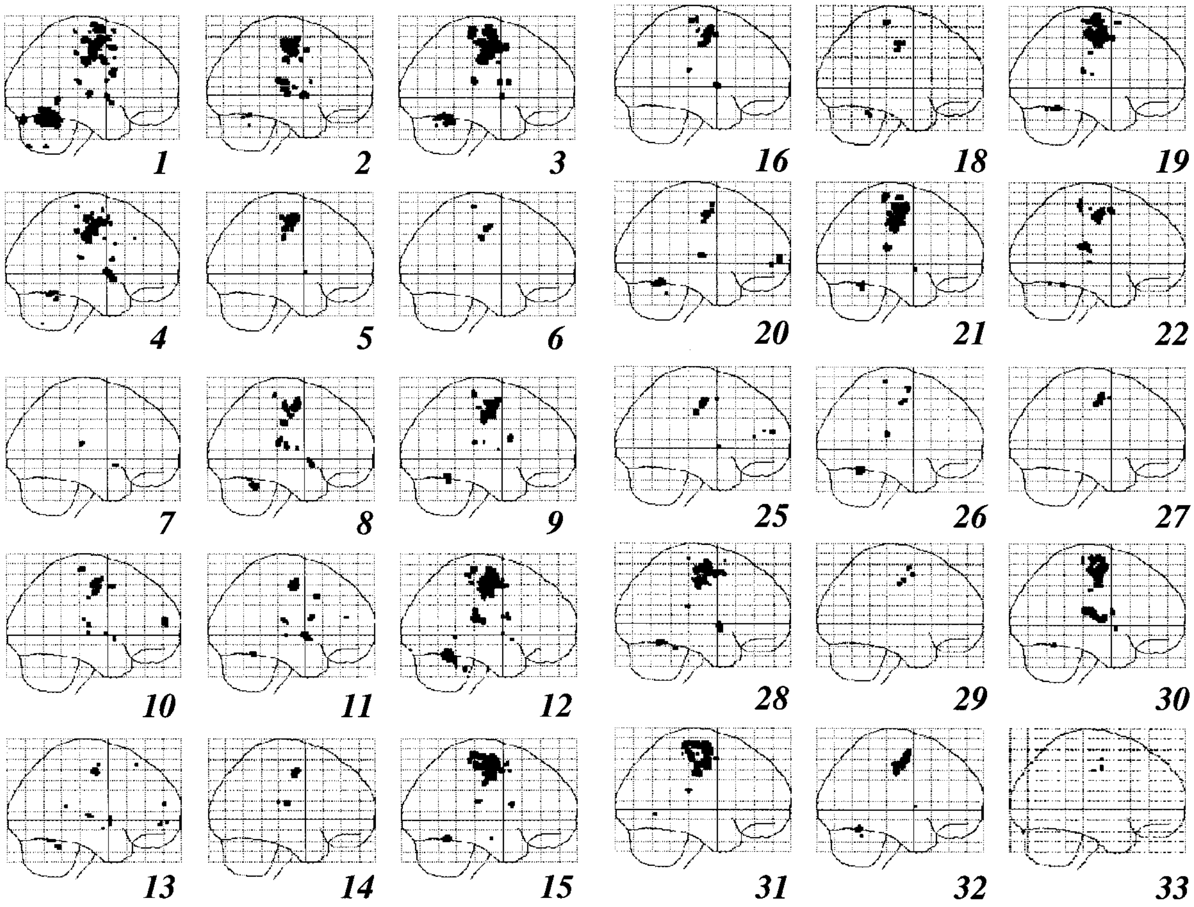

FIG. 2. Single-session sagittal MIPs for the motor paradigm. The number of each session is displayed below it. Although 33 sessions were collected, only 30 are shown here (sessions 17, 23, and 24 were rejected due to movement artifacts). All results are thresholded at $P < 0.05$ corrected for multiple comparisons unless otherwise stated.

This influence of session context on the activation effects of a study constitutes a *session-by-condition* interaction. Although a number of studies have examined the reproducibility of fMRI across a small number of sessions (Cohen *et al.*, 1999; Noll *et al.*, 1997; Rombouts *et al.*, 1998; Tegeler *et al.*, 1999; Yetkin *et al.*, 1996), our primary aim was to examine how well a single session *typifies* a subject's responses. Just as the significance of within-session experimental effects is assessed by sampling a number of scans for each condition, to assess *between-session* differences one must sample multiple sessions. If a single session is to be a good exemplar of a subject's functional neuroanatomy, session-by-condition interactions must be minimal.

The issue of single-session generality also influences data analysis. If activation effects do indeed vary substantially between sessions, to generalize the results to the subject an experiment will need to utilize multiple sessions and assess the data accounting for both within- and between-session variability. Typically, these two levels of variability are not addressed, even if multiple sessions are acquired; the experimental effects of interest are assessed using statistical models that utilize within-session error variance (residual scan/scan variability) as the only component of variance. Although session-by-condition interactions are often modeled, the variability of these interaction effects does not enter into the inference. Such a model, employing a single variance component, is a *fixed-effects* model (Searle *et al.*, 1992). These models have been the norm in neuroimaging analysis and assess only the average experimental effect across the observed sessions. They do not take account of the variability of responses between sessions and therefore cannot be used to draw conclusions about a subject's typical response. For example, a spuriously large activation in one voxel during only one session may be large enough to dominate that voxel's average responses across sessions. In the case of a single session collected from a single subject, the experiment is reduced to a case study. Conclusions regarding the subject's typical response can be made only under the implicit assumption that intersession variability of response would be negligible were the experimental ses-

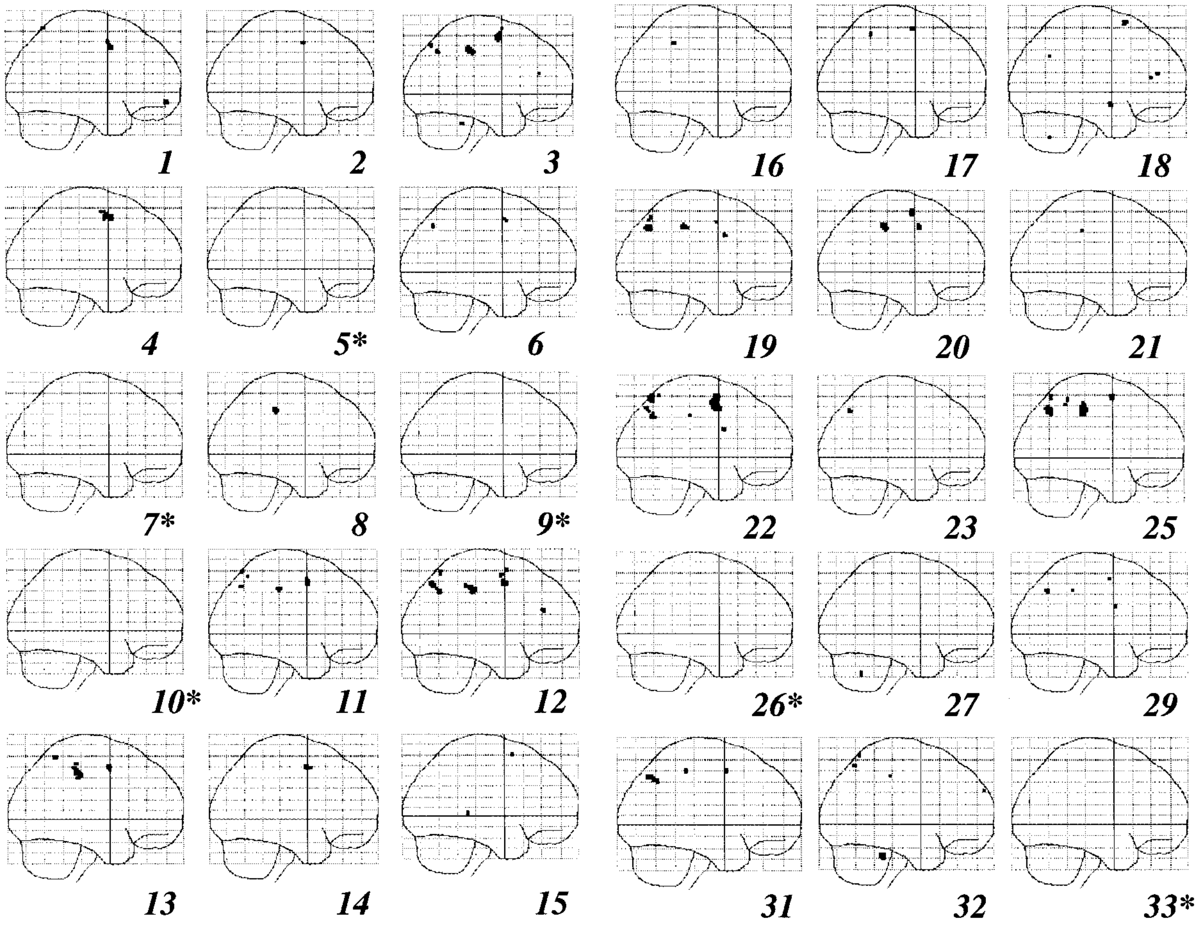

FIG. 3. Single-session sagittal MIPs for the cognitive paradigm. Similar to Fig. 2, although 33 sessions were collected, only 30 are displayed. Sessions marked with "*" contain no significant voxels.

sion repeated. As discussed above, this is highly unlikely.

If session-by-condition interactions are substantial, *random-effects* models are required. Random-effects models allow for multiple variance components (Searle *et al.*, 1992), so the effects of each session on the BOLD response are treated as a random variable. This reflects the fact that a single session is considered a sample from the population of all possible sessions from the subject, and so significance can be computed, accounting for both between- and within-session variance. Random-effects analyses have previously been employed to account for between-subject variability, or *subject-by-condition interactions,* in fMRI studies (Holmes *et al.*, 1998; Henson *et al.*, 1999a,b).

As the random-effect analysis infers about the population from which the samples were drawn, the *N* of observations is now the number of sessions. As the number of sessions is quite small, these analyses tend to have low power, that is, there is a high chance of type II errors. An analysis of this type, however, is essential for the correct level of inference if session-by-condition interactions are considerable. In the present study we examined the reproducibility of the BOLD response in a *single* subject over multiple sessions for simple motor, cognitive, and visual paradigms. We first present results from each session analyzed in isolation, as if from a single-session experiment, using only within-session variance to compute significance. We then show where significant session-by-condition interactions occur for each of our activation paradigms. Finally, we consider two analyses of the entire multisession data set. The first is a fixed-effects analysis, the second a simple random-effects analysis.

METHODS

Subject and Session Details

The subject was a healthy 23-year-old right-handed male. As our goal in the current study was to examine the generality of a single session, each session was conducted as if it were the first time the subject had been examined: in effect, as if only one session was to be obtained. Our motivation was therefore to control for obvious and artifactual between-session differences

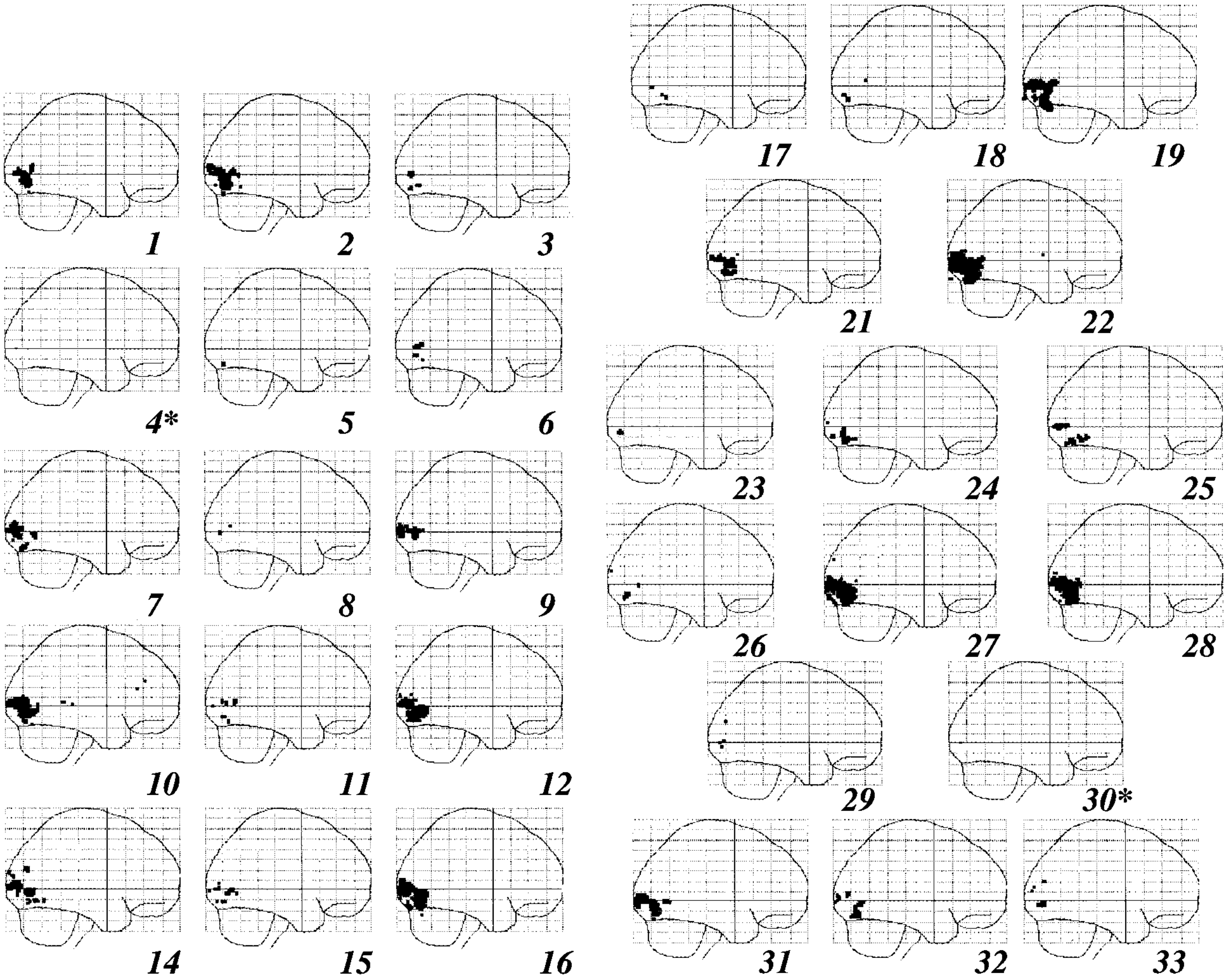

FIG. 4. Single-session sagittal MIPs for the visual paradigm. Similar to Figs. 2 and 3, only 31 sessions are displayed. Sessions marked with "*" contain no significant voxels.

while ensuring that sources of typical between-session variability (scanner hardware and subject physiology) would be sampled in an unbiased manner. The following precautions were taken: the same operators always controlled the scanner, ambient light and sound levels were similar between sessions, and spoken instructions to the subject were always exactly the same. One obvious factor that we could not control was that our subject was always aware that he had performed the task before in the scanner, only under slightly different circumstances. We called this the "Groundhog Day" effect.

Ninety-nine individual sessions were acquired from the subject over a period of 2 months. Each scanning session consisted of one run of a motor, cognitive, or visual paradigm. To minimize scanning time, sessions were acquired in blocks of three. Each block of three sessions comprised a motor, visual, and cognitive session. The order of sessions across scanning blocks was randomized to balance any possible order effects. Session paradigms were designed to reduce the effects of variable task performance. For example, the subject was familiarized with both the random number generation and the finger-tapping task before performing them in the scanner, in an attempt to eliminate performance effects. In addition, the rates at which both tasks were performed were chosen to ensure that subject performance would be stable across sessions. These decisions were informed by studies which used similar paradigms (motor paradigm—Blinkenberg *et al.*, 1996; cognitive paradigm—Jahanshahi *et al.*, submitted for publication).

Motor Paradigm

The subject tapped his right index finger, paced by an auditory tone (1.5 Hz). The subject's hand was restrained within a custom-built thermoplastic splint, which ensured that the amplitude of the finger movement was consistent both across and within sessions.

TABLE 1

			Talairach coordinates		
Area	Cluster size (*k*)[a]	*Z* score[b]	*X*	*Y*	*Z*
(a) Local maxima for the motor fixed-effects model					
Left precentral gyrus (SM1)	5958	9.77	−38	−10	52
Left precentral gyrus (SM1)		9.63	−62	−20	38
Left postcentral gyrus (SM1)		9.51	−60	−18	46
Right cerebellum, anterior lobe	1709	9.50	20	−54	−18
Right cerebellum, anterior lobe		9.19	2	−54	−24
Right cerebellum, anterior lobe		9.14	8	−58	−20
Left supplementary motor area (SMA)	927	9.39	−2	−2	52
Right SMA		8.75	8	2	58
Right SMA		8.50	4	8	66
Left inferior frontal gyrus	227	9.23	62	6	16
Left inferior frontal gyrus		6.98	62	4	4
Right postcentral gyrus (SM1)	197	9.07	58	−12	50
Right precentral gyrus (SM1)		8.29	52	−6	38
Right precentral gyrus (SM1)		7.71	62	−2	38
Left inferior thalamus	466	8.84	−12	−18	2
Left ventral midbrain		6.15	−8	−12	−14
Left midbrain		5.90	−8	−20	−12
Right lateral premotor cortex	70	8.80	38	−8	50
Right inferior frontal gyrus	62	8.60	64	0	16
Right inferior parietal lobule	604	8.54	50	−28	24
Right transverse temporal gyrus		8.36	52	−18	12
Right superior temporal gyrus		7.99	68	−26	16
Left posterior cingulate gyrus	100	8.32	−8	−26	42
Left posterior cingulate gyrus		7.11	−2	−22	48
Right inferior frontal operculum	228	8.23	0	6	4
Right insula		6.12	0	6	10
Left cerebellum, anterior lobe	66	7.85	20	−56	−20
Right cerebellum, anterior lobe	98	7.73	2	−58	−48
Right cerebellar tonsil		7.43	6	−70	−40
Left posterior postcentral gyrus	67	7.63	18	−46	64
Left superior frontal gyrus	79	7.49	28	42	26
Left superior thalamus	46	7.40	18	−10	18
Left thalamus		5.73	−12	−6	12
Left medial frontal gyrus	223	7.27	−38	58	−2
Left medial frontal gyrus		7.27	−30	62	4
Left inferior frontal gyrus		7.20	−34	54	−8
Third ventricle	22	7.27	6	−42	−6
Right putamen	171	7.26	8	6	−2
Right putamen		6.68	0	6	−12
Right anterior cingulate gyrus	27	6.74	14	8	42
Right postcentral gyrus	7	6.74	40	−30	58
Right thalamus	40	6.70	14	−14	2
Right inferior frontal gyrus	18	6.46	30	60	−2
White matter, subcortical	33	6.46	16	0	12
White matter, subcortical		5.13	12	6	2
Right posterior cingulate	19	6.43	12	−24	46
Right cerebellum	18	5.94	48	−50	−38
Right thalamus	16	5.79	20	0	−6
Left cerebellum	12	5.76	−36	−64	−26
Left superior frontal gyrus	5	5.64	−28	28	56
Right thalamus	9	5.55	8	−4	−4
Right cerebellum	12	5.51	40	−62	−50
Left superior frontal gyrus	8	5.42	−22	38	48
Right caudate nucleus	9	5.30	20	−6	20
(b) Local maxima for the motor random-effects analysis					
Left precentral gyrus (SM1)	4253	8.76	−36	−10	52
Left precentral gyrus (SM1)		8.45	−60	−18	38
Left frontal operculum		8.37	−48	4	0
Right cerebellum, anterior lobe	951	8.04	18	−52	−18
Right cerebellum, anterior lobe		7.55	30	−52	−26
Right cerebellum, anterior lobe		7.34	36	−62	−24

TABLE 1—*Continued*

Area	Cluster size (k)[a]	Z score[b]	Talairach coordinates		
			X	Y	Z
Left inferior frontal gyrus	151	8.00	−62	8	16
Left inferior precentral gyrus		6.96	−60	−2	16
Left supplementary motor area (SMA)	545	7.62	−8	−2	48
SMA		7.34	0	2	56
Right superior SMA		7.15	4	10	66
Left thalamus	173	7.44	−12	−20	2
Left midbrain (red nucleus)		5.47	−12	−14	−10
Right postcentral gyrus (SM1)	104	7.40	58	−12	50
Right inferior precentral gyrus		6.93	50	−4	36
Right inferior precentral gyrus		5.75	60	−2	40
Left supramarginal gyrus	44	7.27	−58	−44	18
Right lateral premotor cortex	46	7.19	38	−8	52
Right frontal operculum	104	7.10	50	6	4
Right parietal operculum	256	7.07	50	−28	26
Right transverse temporal gyrus		6.94	48	−18	14
Right transverse temporal gyrus		6.16	68	−26	16
Left posterior cingulate gyrus	39	6.88	−8	−24	44
Right inferior frontal gyrus	32	6.71	64	0	16
Right inferior postcentral gyrus	25	6.56	62	−12	12
Cerebellar vermis	25	6.22	6	−70	−40
Left superior frontal gyrus	21	6.04	−28	42	26
Right cerebellar tonsil	8	5.88	12	−56	−48
Left superior thalamus	12	5.81	−20	−12	18
Left posterior postcentral gyrus	14	5.73	−18	−46	64
Left cerebellum, anterior lobe	10	5.68	−20	−56	−20
Right putamen	20	5.57	28	8	−2

Note. SM1, primary somatomotor cortex; SMA, supplementary motor cortex.

[a] Only clusters with $k \geq 5$ are listed. Where multiple foci exist for a cluster, the three most significant are reported.

[b] All foci reported survive a statistical threshold of $P < 0.05$ corrected for multiple comparisons.

Each activation epoch was alternated with a rest epoch, in which the pacing tone was delivered to control for auditory activation. Thirteen blocks were collected per session (7 rest and 6 active). The subject maintained fixation on a cross that was backprojected onto a transparent screen by an LCD video projector. The projector was similarly employed to deliver visual instructions to the subject before each block (either “Move” or “Rest”).

Cognitive Paradigm

The subject generated random numbers from 1 to 9, paced by an auditory tone (0.66 Hz). In the rest condition the subject counted from 1 to 9, similarly paced by the auditory tone. The subject fixated in a fashion similar to that used before. Thirteen epochs were collected in total (7 rest and 6 active).

Visual Paradigm

A reversing black and white checkerboard flickering at 8 Hz (Fox and Raichle, 1985) was presented to the subject. The subject focused on a central fixation spot that was constant across both activation (reversing checkerboard stimulation) and rest (fixation spot only) blocks. Six epochs were acquired in total (three activation and three rest).

Scanning Parameters

The data were acquired on a Siemens Magnetom Vision (Siemens, Erlangen, Germany) at 2 T. Each BOLD–EPI volume scan consisted of 48 transverse slices (in-plane matrix 64 × 64; voxel size 3 × 3 × 3 mm; TE = 40 ms; TR = 4.1 s). Seventy-eight volume scans were collected during each cognitive and motor session, and 36 scans per visual session (epoch length was always 6 scans). A T1-weighted high-resolution MRI of the subject (1 × 1 × 1.5 mm resolution) was acquired to facilitate anatomical localization of the functional data.

Image Preprocessing

Data preprocessing was carried out using SPM99 (Wellcome Department of Cognitive Neurology, London, UK; http:/www.fil.ion.ucl.ac.uk/spm). All functional volumes, independent of session or paradigm, were realigned to the first volume acquired (Friston *et al.*, 1995) and a mean realigned volume was created. Sessions containing obvious movement artifacts (de-

TABLE 2

			Talairach coordinates		
Area	Cluster size (k)[a]	Z score[b]	X	Y	Z
(a) Local maxima for the cognitive fixed-effects analysis					
Right inferior parietal lobule	888	9.27	50	−30	44
Right inferior parietal lobule		8.63	34	−44	44
Right supramarginal gyrus		7.95	66	−24	40
Left medial precentral gyrus (FEF)	1716	9.26	−24	0	50
Left lateral precentral gyrus		8.96	−50	4	46
Left middle frontal gyrus		8.51	−54	10	26
Left precuneus	1812	9.17	−10	−66	44
Left superior parietal lobule		8.86	−42	−36	50
Left superior parietal lobule		8.77	−18	−62	64
Right medial precentral gyrus (FEF)	776	9.17	26	−4	58
Right superior frontal gyrus		8.70	20	6	60
Right superior frontal gyrus		7.68	12	14	66
Right superior parietal lobule	876	8.99	20	−66	58
Right superior parietal lobule		8.74	26	−70	50
Right superior parietal lobule		8.42	22	−58	62
Right anterior precentral gyrus	238	8.84	62	4	28
Left middle frontal gyrus	454	8.63	−34	36	22
Left middle frontal gyrus		8.22	−48	40	24
Left middle frontal gyrus		7.11	−30	46	34
Right lateral premotor cortex	85	8.39	56	0	46
Left supplementary motor area	421	8.30	−2	18	50
Right inferior frontal gyrus	406	8.30	56	16	−2
Right inferior frontal gyrus		7.44	48	26	−8
Right superior temporal gyrus		7.44	−34	16	6
Left inferior frontal gyrus	414	8.24	−56	12	0
Left inferior frontal gyrus		8.09	−42	22	−8
Left insula		7.23	−34	16	6
Right cerebellum, anterior lobe	190	8.00	30	−62	−30
Right cerebellum, anterior lobe		7.32	34	−54	−36
Right cerebellum, anterior lobe		5.86	40	−60	−36
Left tempero-occipital sulcus	41	7.96	−52	−52	−18
Right inferior parietal lobule	44	7.92	68	−38	24
Right inferior parietal lobule		5.48	68	−30	32
Right inferior parietal lobule		5.33	64	−44	34
Right middle frontal gyrus	311	7.91	44	40	32
Right middle frontal gyrus		7.81	44	42	20
Right middle frontal gyrus		7.80	34	36	18
Right anterior cingulate gyrus	37	7.75	12	12	38
Left inferior parietal lobule	80	7.67	−62	−36	32
Right middle temporal gyrus	72	7.65	58	−20	−6
Right superior temporal gyrus	67	7.62	48	−32	−2
Right middle occipital gyrus	27	7.21	50	−74	12
Left precentral gyrus	11	6.99	−54	−8	14
Left middle frontal gyrus	28	6.99	−30	56	26
Left superior temporal gyrus	43	6.96	−54	−40	4
Left inferior frontal gyrus	16	6.67	−54	34	8
Right cerebellum	16	6.58	36	−38	−42
Right calcarine cortex (V1)	17	6.45	12	−78	8
Right insula	11	6.39	40	6	0
Left cerebellum	30	6.36	−22	−56	−34
Left anterior cingulate	9	6.35	−12	16	32
Left inferior frontal gyrus	23	6.02	−36	58	−2
Left inferior frontal gyrus		5.53	−32	62	4
Right insula	11	5.90	42	16	4
Left calcarine cortex (V1)	17	5.80	−6	−72	20
Right middle frontal gyrus	7	5.77	40	30	24
Left calcarine cortex (V1)	6	5.50	−4	−9	0
Left hemisphere, white matter.	9	5.44	−22	54	0
Right superior frontal gyrus	5	5.24	16	38	54

TABLE 2—*Continued*

Area	Cluster size (k)[a]	Z score[b]	Talairach coordinates		
			X	Y	Z
(b) Local maxima for the cognitive random-effects analysis					
Left superior parietal lobule	690	8.13	−40	−36	48
Left posterior postcentral gyrus		7.44	−48	−28	40
Left superior parietal lobule		7.28	−30	−48	42
Left medial precentral gyrus (FEF)	1134	8.07	−26	−2	50
Left middle frontal gyrus		8.06	−50	8	44
Left ventral precentral gyrus		7.74	−52	10	26
Left superior parietal lobule	156	7.87	−18	−64	64
Left superior parietal lobule		6.58	−26	−62	60
Left superior parietal gyrus		5.59	−34	−60	60
Right inferior parietal lobule	460	7.87	48	−30	44
Right inferior parietal lobule		7.50	56	−32	44
Right supramarginal gyrus		6.84	64	−24	40
Right superior parietal lobule	540	7.75	18	−64	58
Right superior parietal lobule		6.97	22	−56	62
Right superior parietal lobule		6.93	26	−64	34
Right medial precentral gyrus (FEF)	569	7.73	28	−2	60
Right superior frontal gyrus		7.62	20	6	62
Right superior frontal gyrus		6.47	12	12	68
Left precuneus	157	7.72	−8	−66	42
Right anterior precentral gyrus	173	7.56	62	6	28
Left middle frontal gyrus	276	7.38	−34	38	24
Left middle frontal gyrus		6.81	−50	36	24
Right cerebellum, anterior lobe	78	7.24	30	−62	−30
Right cerebellum, anterior lobe		6.56	32	−56	−36
Right inferior parietal lobule	22	7.03	68	−38	24
Right middle temporal gyrus	55	6.95	60	−20	−8
Right lateral premotor cortex	41	6.94	54	2	44
Left inferior frontal gyrus	139	6.92	−42	22	−8
Left inferior frontal gyrus		6.80	−56	12	0
Right inferior frontal gyrus	122	6.89	56	16	−2
Left temporal-occipital sulcus	18	6.77	−52	−52	−18
Left supplementary motor area (SMA)	201	6.71	−2	18	48
Left supplementary motor area (SMA)		6.66	−6	10	54
Left supplementary motor area (SMA)		6.64	0	16	58
Right anterior cingulate gyrus	15	6.66	14	12	38
Right middle frontal gyrus	98	6.53	38	38	34
Right middle frontal gyrus		5.93	44	40	20
Right middle frontal gyrus		5.66	34	44	38
Left inferior parietal lobule	33	6.26	−64	−38	26
Right middle frontal gyrus	19	6.26	34	36	18
Right superior temporal gyrus	26	6.01	46	−30	−4
Left middle frontal gyrus	7	5.98	−30	54	28
Right superior temporal gyrus	16	5.90	54	18	−20
Right inferior frontal gyrus	16	5.79	50	26	−8
Left superior frontal gyrus	7	5.73	−10	12	70

Note. FEF, frontal eye fields; V1, primary visual cortex; SMA, supplementary motor area.

[a,b] See footnotes to Table 1.

cided by two of the authors with previous experience of typical artifacts) were discarded at this stage: three motor sessions, two visual sessions, and three cognitive sessions were excluded in this manner. The subject's T1-weighted structural scan was coregistered to the mean functional volume, and the mean volume used to determine the parameters applied to all volumes during spatial normalization and resampling (Ashburner *et al.*, 1997; Ashburner and Friston, 1999) to a standard template (Evans *et al.*, 1993). As the volume of brain sampled in each study was affected by the position of the subject within the scanner's field of view, we found that the extreme superior and inferior portions of the subject's brain were sparsely sampled. To address this, voxels not sampled in *every* session were eliminated during normalization. All functional volumes were then smoothed with a FWHM Gaussian kernel. Global changes in fMRI response from scan to

TABLE 3

			Talairach coordinates		
Area	Cluster size (*k*)	*Z* score	*X*	*Y*	*Z*
(a) Local maxima for the visua fixed-effects analysis					
Right calcarine cortex (V1)	11002	9.72	14	−86	2
Right calcarine cortex (V1)		9.63	6	−74	−4
Left calcarine cortex (V1)		9.57	−8	−82	0
Right superior temporal gyrus	80	7.96	42	−28	18
Right precuneus	54	7.93	8	−80	44
Left lateral geniculate nucleus	91	7.92	−18	−26	−4
Right superior temporal gyrus	51	7.81	70	−32	14
Left inferior parietal lobule	152	7.79	−32	−38	54
Left superior parietal lobule	130	7.75	−28	−54	54
Left superior parietal lobule		6.66	−20	−52	54
Left superior parietal lobule		5.08	−28	−48	62
Left superior temporal gyrus	32	7.60	−56	−42	16
Right lateral geniculate nucleus	83	7.59	24	−24	4
Right postcentral gyrus	95	7.52	36	−32	54
Left superior temporal gyrus	31	7.17	−42	−28	20
Right lateral ventricle	112	6.82	20	−26	24
Left precuneus	25	6.76	−18	−62	48
Right midbrain	27	6.71	8	−26	−6
Right middle frontal gyrus	22	6.58	60	2	38
Right inferior frontal gyrus	26	6.46	56	36	12
Right inferior frontal gyrus		5.50	52	42	16
Left lateral ventricle	58	6.30	−16	−22	26
Right precuneus	12	6.24	28	−62	48
Right cerebellum	18	6.18	12	−70	−46
Right lateral ventricle	11	6.18	4	12	12
Right parietal lobe, white matter	7	6.18	36	−58	64
Right supplementary motor area	8	5.50	2	2	64
Right cerebellum	6	5.49	32	−72	−30
Right fusiform gyrus	5	5.48	48	−50	−16
Right temporal lobe, white matter	6	5.40	34	−40	10
Right superior parietal lobule	5	5.20	40	−52	54
(b) Local maxima for the visual random-effects analysis					
Right calcarine cortex (V1)	8766	8.66	10	−80	−8
Right extrastriate cortex		8.59	24	−96	16
Right extrastriate cortex		8.54	22	−76	−18
Right precuneus	46	7.48	8	−78	44
Right parieto-occipital fissure		7.23	12	−86	38
Right superior temporal gyrus	80	6.68	42	−28	18
Left superior parietal lobule	37	6.49	−26	−56	54
Left lateral geniculate nucleus	27	6.21	−16	−28	−4
Left inferior parietal lobule	47	6.06	−32	−38	56
Right postcentral gyrus	17	6.04	34	−32	52
Right lateral geniculate nucleus	21	5.98	22	−26	0
Right superior temporal gyrus	17	5.96	70	−34	12
Right inferior precentral gyrus	6	5.87	64	10	6
Left intraparietal sulcus	13	5.86	−20	−70	34
Left superior temporal gyrus	9	5.72	−40	−26	20
Right intraparietal sulcus	26	5.71	32	−70	26
Left superior temporal gyrus	7	5.70	−56	−42	18
Right middle frontal gyrus	6	5.57	58	4	38
Right intraparietal sulcus	8	5.34	24	−76	38
Left inferior occipital gyrus	5	5.16	−40	−64	−8

Note. V1, primary visual cortex.

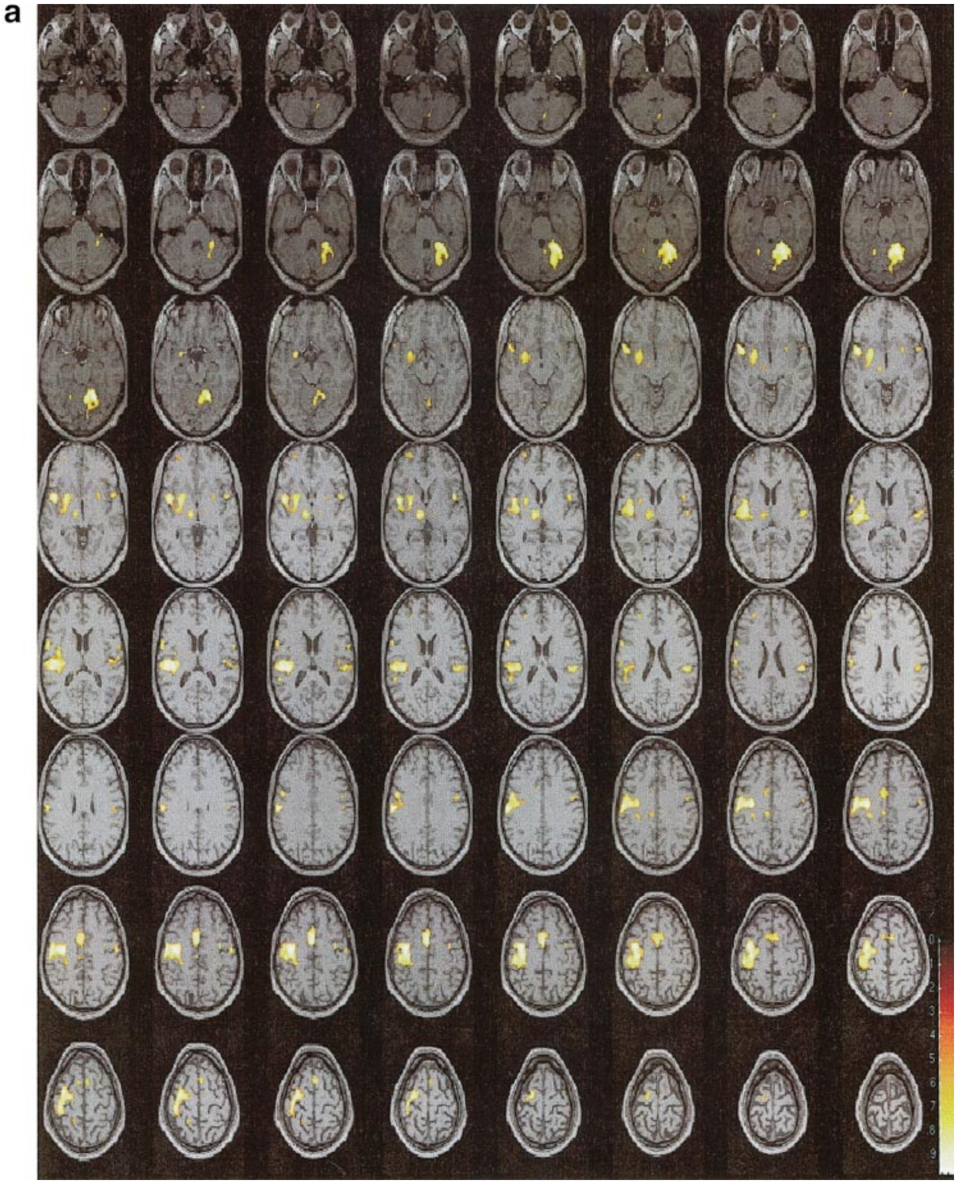

FIG. 5. Multisession analyses of the motor paradigm, analyzed using a fixed-effects model (a), extra-sum of squares F test (b), and a random-effects model (c). Voxels surviving the statistical threshold are displayed on a coregistered structural scan of the subject to aid the identification of activated areas. Each transverse slice is 2 mm thick. The color bar represents statistical significance, with higher Z and F scores having a brighter color.

scan were removed by proportionally scaling each scan to have a common global mean voxel value.

Theory and Implementation

Statistical analysis was carried out using the general linear framework described by Worsley and Friston (1995). The sessions for each paradigm were modeled with a simple linear model for the data at each voxel:

$$Y_{ij} = \gamma_i + \alpha_i \mathrm{f}(j) + \sum_{k=1}^{K} \beta_{ik} g_k(j) + \epsilon_{ij}. \qquad (1)$$

Here Y_{ij} denotes the value of the voxel of scan j ($j = 1, \ldots, J$) of session i ($i = 1, \ldots, I$). γ_i is the mean (block) effect for session i. $\mathrm{f}(j)$ is a reference waveform, a function of the scan index within session which has the same form for all sessions. Here we shall use a simple "convolved boxcar" reference waveform (CBC), consisting of a boxcar function of zeros and ones representing the experimental time course, convolved with the expected hemodynamic response function. The parameter α_i is the amplitude of the CBC response for session i. Differences in the session response amplitudes α_i constitute session-by-condition interac-

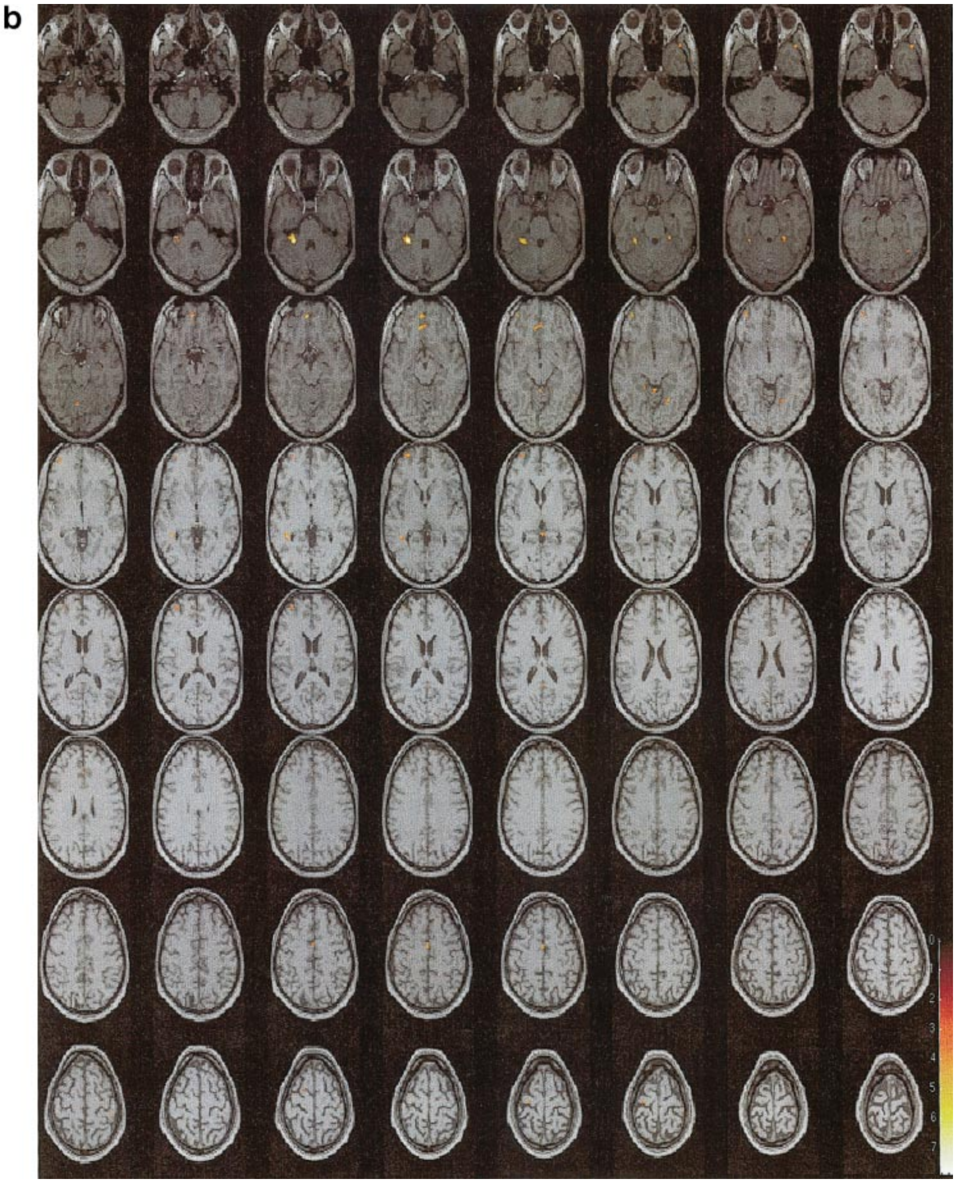

FIG. 5—*Continued*

tions. The additional reference functions $g_k(j)$ are a set of discrete cosine basis functions, effecting a simple "high-pass" filter, as described by Holmes *et al.* (1997), with cutoff (specified by K) set at twice the experimental period. We shall assume that this model fits, such that the residual errors (ϵ_{ij}) have zero mean and exhibit only short-term autocorrelation within session. In the following we shall refer to the CBC amplitudes α_i simply as the response for session i.

Individual-Session Analyses

Each session was analyzed alone as a single fMRI session, as if it were the only session acquired, using a "standard" SPM analysis. The Groundhog Day effects aside, this enables a comparison of how the results of a single-session experiment can vary and illustrates why drawing conclusions about a subject from a single session can be dangerous. The model used is that of Eq. (1), but considering only a single session (i) at a time. The residual errors are assumed to be Normally distributed with variance $\sigma^2_{i(\epsilon)}$, estimated individually for each session. Temporal autocorrelation was dealt with using the method of Worsley and Friston (1995) by temporally smoothing the session time series with a Gaussian kernel of 6-s FWHM. The design matrix for each session is illustrated in Fig. 1A. A t statistic assessing the null hypothesis of zero response ($\alpha_i = 0$) was constructed for each voxel, giving an SPM{t} for

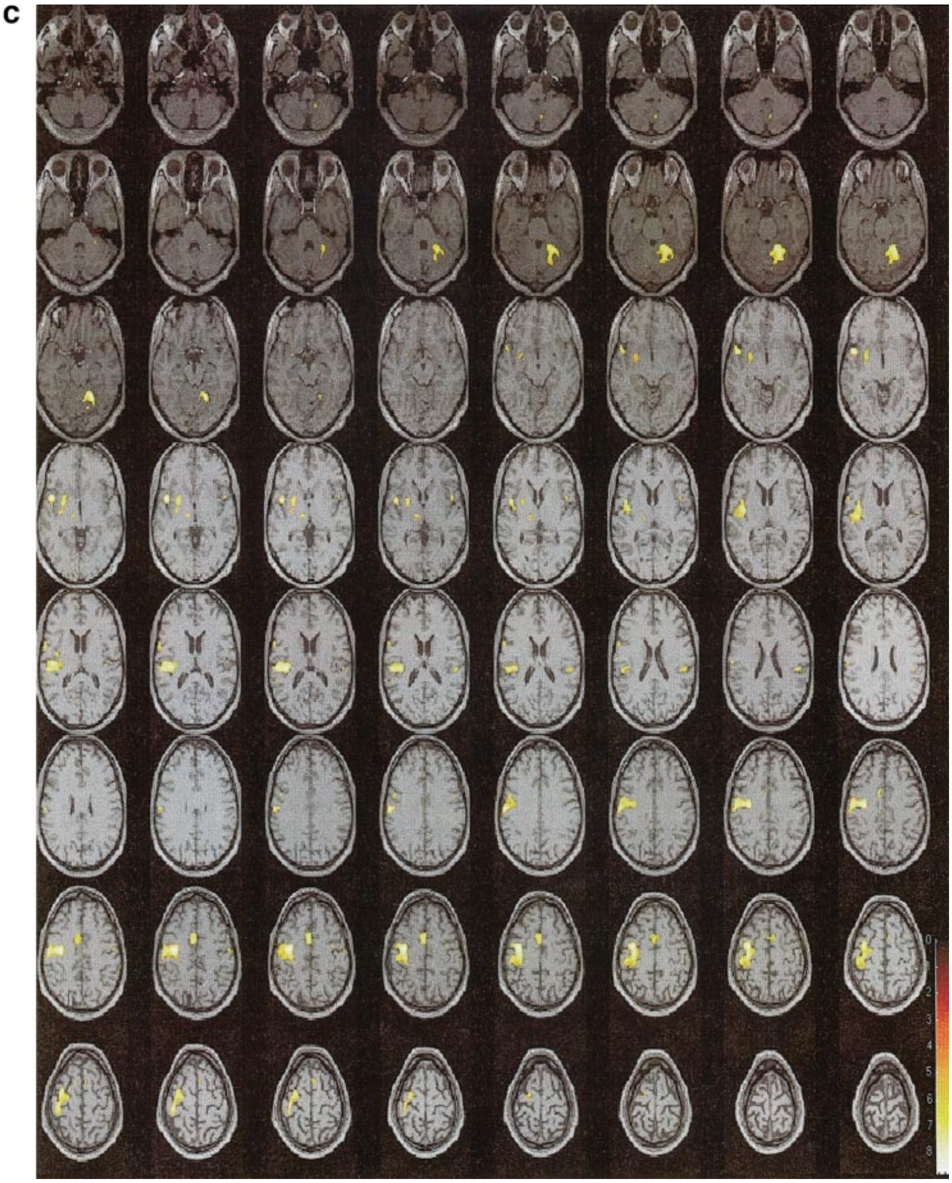

FIG. 5—*Continued*

each session indicating the significance of the response. For display, each session-specific SPM{t} was transformed to an equivalent SPM{Z} by probability integral transform. This was effected by replacing each t value with the standard Normal ordinate with the same upper tail probability.

Multiple-Session Analyses—Session-by-Condition Interactions

To assess whether there were significant session-by-condition interactions, we compared the model of Eq. (1) for all I sessions (design matrix shown in Fig. 1B) with a reduced model in which the response was identical for all sessions ($\alpha_i = \alpha', i = 1, \ldots, I$). Here we assume that the residual variance is identical across sessions, such that the residuals are Normally distributed with zero mean and variance σ_ϵ^2. The additional variance modeled by the full model (including session-by-condition interactions) was compared with the residual variance using an extra sum-of-squares F test (Draper and Smith, 1981), modified to account for temporally autocorrelated residuals using the method of Worsley and Friston (1995). The resulting SPM{F} identifies voxels that display significant session-by-condition interactions.

Multiple-Session Analyses

If there are substantial differences in response from session to session a single-session experiment is inadequate if one wishes to examine a subject's response to experimental stimuli in general, and so a multiple-session experiment is necessitated.

Multiple-Session Analyses—Fixed-Effects Model

Given a multiple-session data set, modeled with Eq. (1) (design matrix shown in Fig. 1B), a fixed-effects analysis proceeds by assuming that the session-specific responses α_i themselves are of interest. The residual errors ϵ_{ij} are assumed Normally distributed with zero mean and constant variance σ_ϵ^2. Evidence of a response across sessions can be tested by examining $\bar{\alpha}.$, the average of the I session-specific responses

$$\bar{\alpha}. = \sum_{i=1}^{I} a_i.$$

Again, short-term temporal autocorrelation in the errors were handled using the method of Worsley and Friston (1995), temporally smoothing each session time series with a Gaussian kernel of 6-s FWHM.

However, since the session-specific responses are considered fixed, only one component of variance is accounted for (the residual error variance σ_ϵ^2), and inference from the resulting SPM{t} is limited to the average response *for the observed sessions.* As such, this analysis is sensitive to large effects in a small number of sessions.

Multiple-Session Analyses—Random-Effects Model

To extend inference beyond the particular sessions acquired, we must recognize that these sessions are merely a sample of possible sessions, each of which would have its own response α_i. Thus, we regard the α_i of Eq. (1) as *random effects,* accepting that the response amplitudes α_i for the sessions under consideration are merely one sample from the (hypothetical) distribution of response amplitudes for a session chosen at random. A simple second-level (between-session) model would be

$$\alpha_i = \alpha + \epsilon_i, \qquad (2)$$

where the α_i are from Eq. (1) (the within-session model), and the between-session errors ϵ_i have zero mean and variance σ_α^2 and can be considered independent. Thus, the random-effects model has two components of variance, between session, σ_α^2, and within session (residual), σ_ϵ^2. Using this model we can consider inference regarding α, the underlying average response across all *possible* sessions.

In general, analysis of such random-effects models can be difficult (Searle and Casella, 1992). However, the simple models considered here are balanced (the models for each session are exactly the same) and separable (the only common parameter across sessions is the intrasession (residual) variance σ_ϵ^2, assumed constant for all sessions). This permits a simple "summary statistic" approach (Frison and Pocock, 1992). Such an approach was first described for neuroimaging data by Worsley *et al.* (1992), and its importance subsequently highlighted by Holmes *et al.* (1998), who describe the implementation (in SPM) used here. In essence, the model of Eq. (1) is fitted to yield estimates $\hat{\alpha}_i$ of the response amplitude α_i at each voxel for each session. The variance of the estimated response amplitudes $\hat{\alpha}_i$ across sessions incorporates both within- (σ_ϵ^2) and between-session variability (σ_α^2) in the appropriate proportions to assess the significance of the overall subject activation effect α (Frison and Pocock, 1992). Thus, each session data set is surmised by a single *contrast* image whose voxel values are the fitted response amplitudes. These contrast images can then be assessed at the intersession level for a significant average effect, with inference extending to the subject in general (under similar experimental conditions) rather than just the particular sessions acquired.

To conduct a parametric analysis, it remains to choose a specific model for the between-session errors ϵ_i. In the absence of any evidence (yet) to suggest otherwise, consider a simple Normal model

$$\alpha_i = \alpha + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\alpha^2). \qquad (3)$$

Our approach here is pragmatic: we know nothing about ϵ_i's distribution. Our assumption of Normality allows us to introduce random effects analyses simply and logically as an extension of the parametric statistical tests used by SPM. We will discuss the validity of this assumption in the discussion.

With the models of Eqs. (1) and (3), the random-effects analysis can be effected as a simple one-sample t test on the contrast images, yielding an SPM{t}.

RESULTS

All tabular data referenced in the results section can be accessed at http://www.fil.ion.ucl.ac.uk/~davem. The coordinate system used is that of the space defined by Talairach and Tournoux (1988).

Individual Session Results

Figures 2, 3, and 4 show sagittal maximum intensity projections (MIPs) per session for the motor, cognitive, and visual tasks, respectively. Each SPM{Z} MIP shows voxels that survive a threshold of $P < 0.05$, corrected for multiple comparisons.

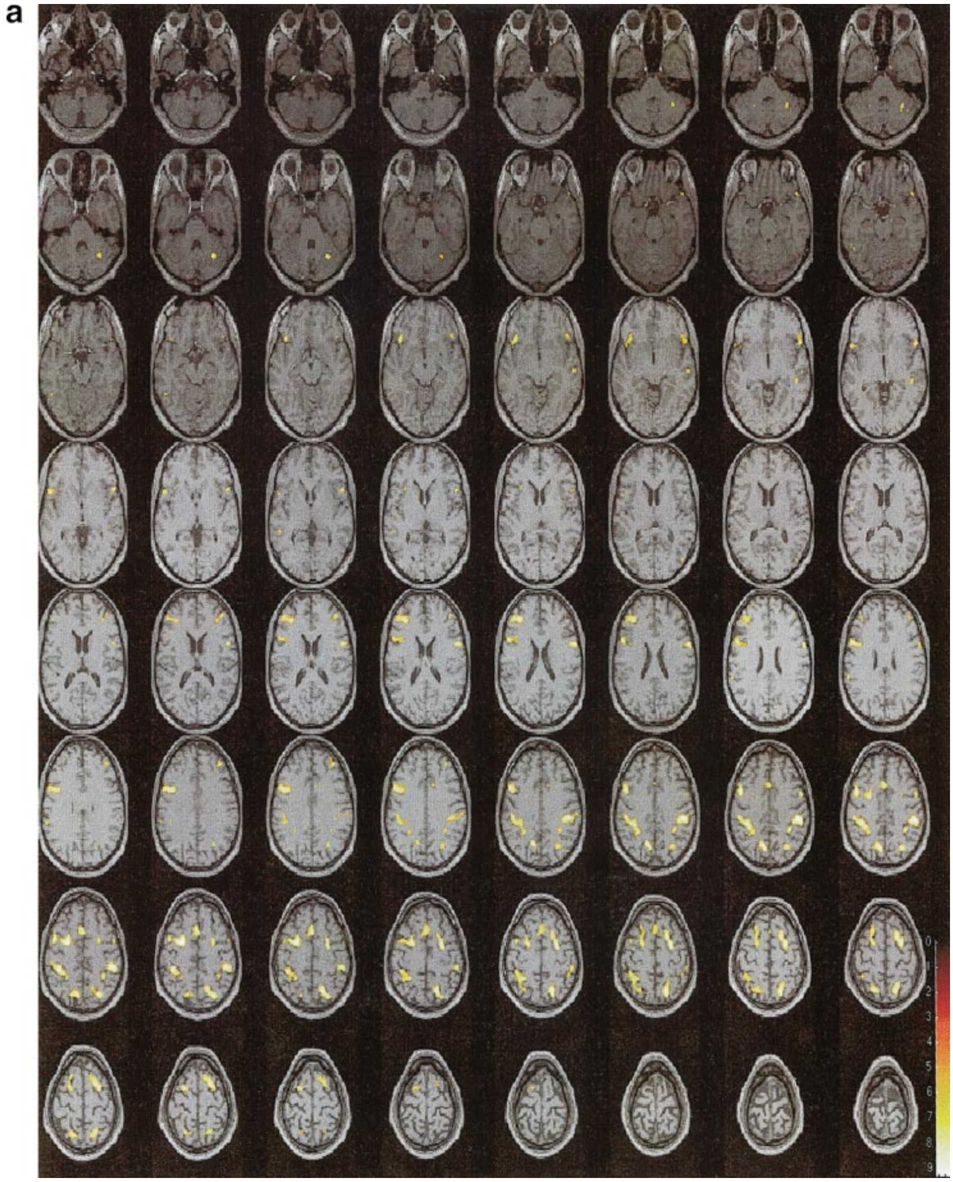

FIG. 6. Multisession analyses of the cognitive paradigm, analyzed using a fixed-effects model (a), extra-sum of squares *F* test (b), and a random-effects model (c). Voxels surviving the statistical threshold are displayed on a coregistered structural scan of the subject to aid the identification of activated areas. Each transverse slice is 2 mm thick. The color bar represents statistical significance, with higher *Z* and *F* scores having a brighter color.

It is immediately obvious that the pattern of activated voxels varies widely between repeated single sessions in our subject. While a grossly homogeneous pattern is evident across single-session MIPs of the same paradigm, the spatial distribution of voxels in each MIP is highly variable. Even though striking similarity is evident between certain data sets (e.g., visual sessions 10 and 12, Fig. 4), a large number of sessions from all three paradigms display *no* significantly activated voxels (e.g., visual sessions 4 and 30). The differences are best exemplified by comparing the SPM{*Z*} of motor session 1 (Fig. 2), which contains 1076 voxels above threshold, and motor session 33, which contains only 5. Results from the cognitive paradigm (Fig. 3) are broadly similar: while the spatial distribution of voxels between MIPs is more comparable than in the motor and visual paradigms, a large number of sessions contain no significantly activated voxels at the chosen threshold.

MIPs are binary statistical images, in which voxels are classified as “active” or “inactive” according to accepted but arbitrary statistical thresholds (for discussions of this issue, see Poline *et al.*, 1996; Genovese *et al.*, 1997; Noll *et al.*, 1997; Cohen and DuBois, 1999; Tegeler *et al.*, 1999). In any of the MIPs of Figs. 2, 3, and 4, a voxel *i* could have very different α_i's between

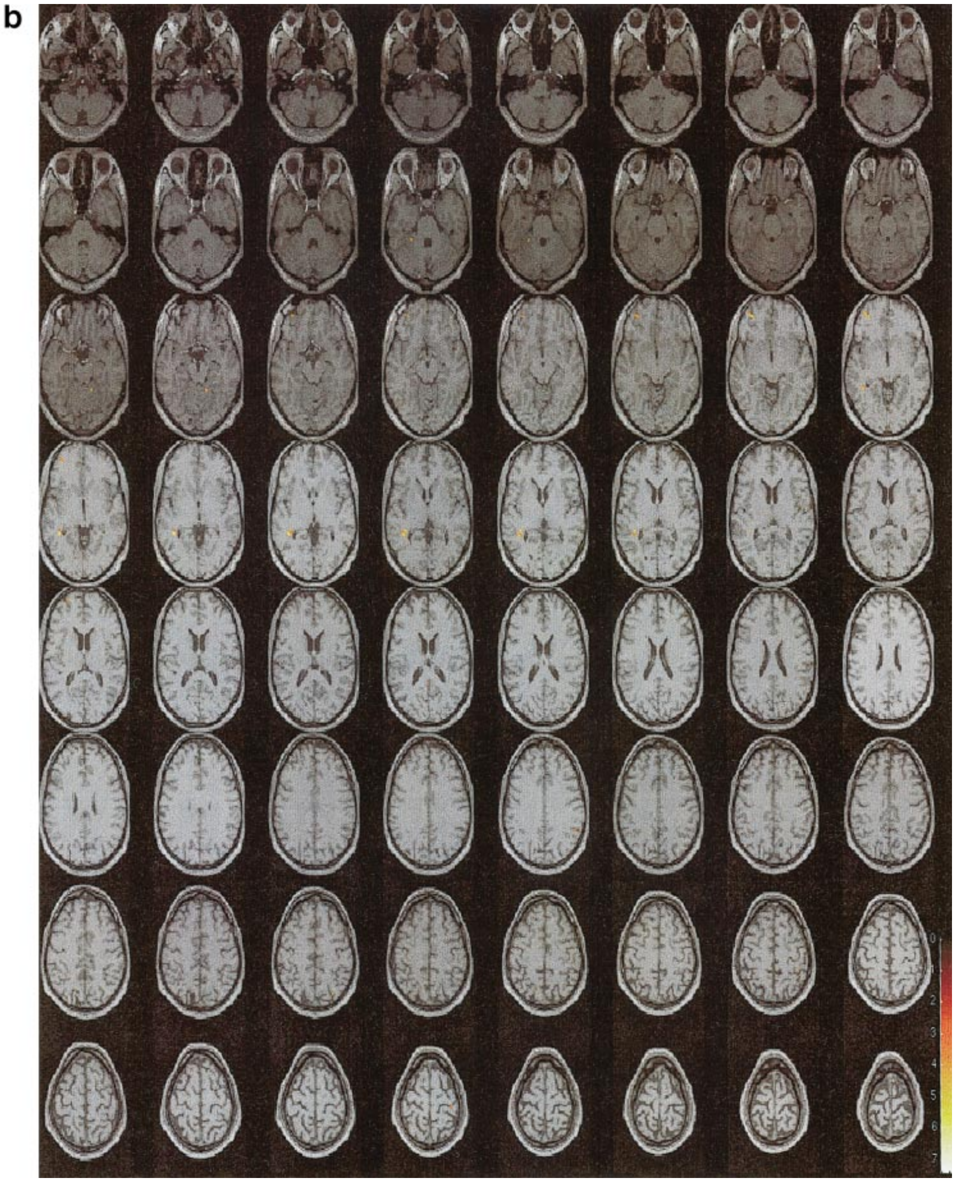

FIG. 6—*Continued*

sessions, yet still pass the threshold and appear to be consistently activated.

Multiple-Session Analyses

Figures 5, 6, and 7 show the results of the motor, cognitive, and visual multiple-session analyses, respectively. As noted above, merely examining thresholded statistical maps is perhaps not the best way to examine similarities between sessions. Our use of the ESS-F test allowed us to examine which voxels showed *statistically significant* variability across all sessions for our single subject (Figs. 5b, 6b, and 7b). If a single session typifies our subject's response, there should be few session-by-condition interactions, and thus the SPM$\{F\}$ maps from each analysis should display relatively few voxels.

By specifically examining the variability of session-by-condition interactions, we implicitly limit our analysis to voxels that are activated on at least one session by the task. Noise that has a truly random expression over time is unlikely to be modeled sufficiently well by each session's regressor of interest; however, task-correlated noise, such as movement, will still present a problem.

ESS{F} Analyses

Figures 5b, 6b, and 7b show the results of each multisession ESS-F test. These SPMs were thresholded at $P < 0.05$ corrected as for the fixed-effects SPM$\{Z\}$'s, reflecting that we did not have any *a priori* hypotheses concerning where we expected to see greater variability. An important point to note at this

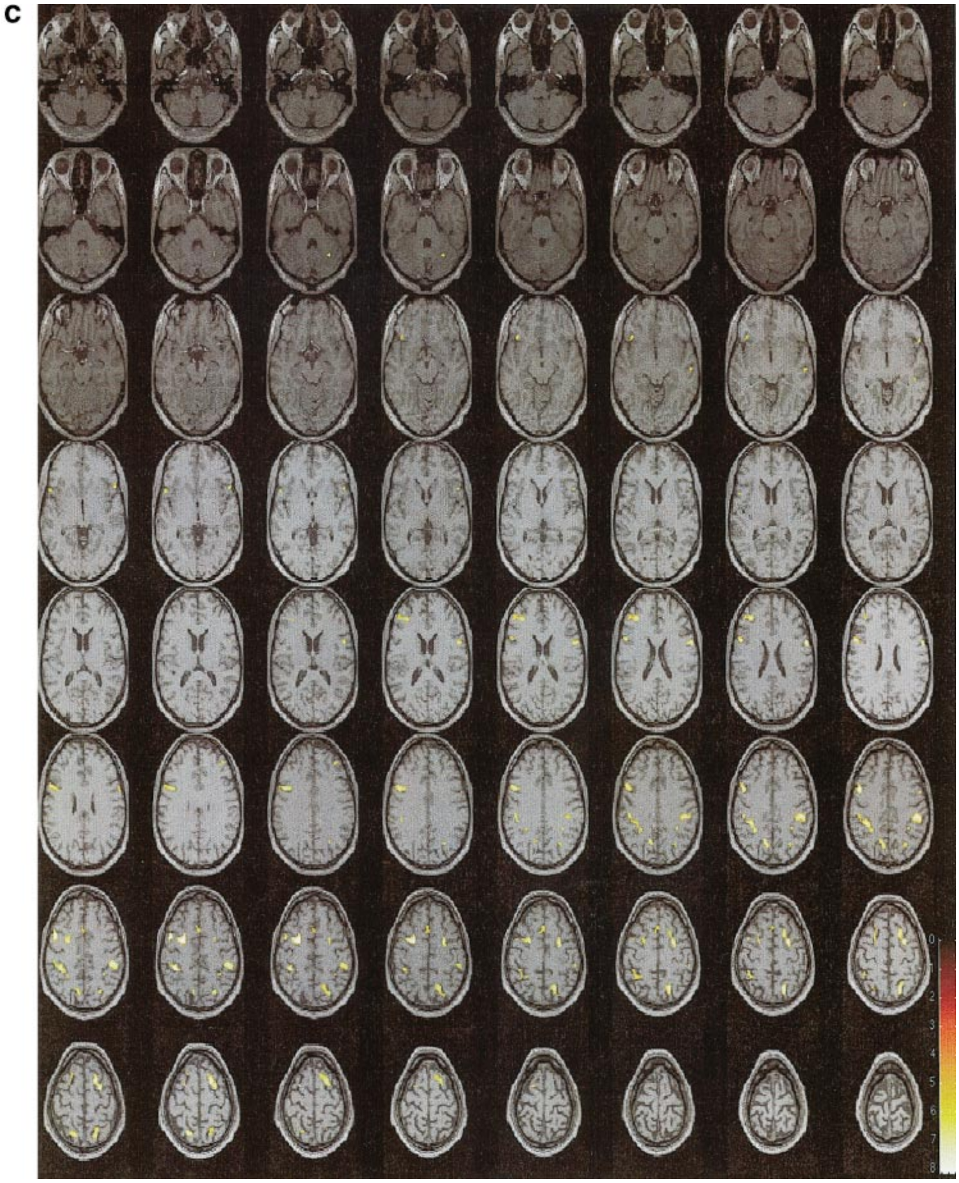

FIG. 6—*Continued*

stage is that the ESS-F test is free of any constraints about the direction of activation effects observed. As such, although our main concern was to examine the variability of activation effects, each SPM$\{F\}$ also contains voxels that had highly variable *deactivations*. In the interest of parsimony, these results will not be discussed here.

Somewhat surprisingly, each fixed-effects SPM$\{Z\}$ did not display a high degree of overlap with its corresponding SPM$\{F\}$. This is because each SPM$\{F\}$ identifies voxels that show high variability, even if they are not classified as activated on average. Reflecting this fact, the area displaying the highest degree of variability in signal intensity between sessions in the motor paradigm (Fig. 5b) is located within the white matter of the temporal lobe (−28, −42, −28, F = 7.88)—an area which does not appear on the fixed-effect SPM$\{Z\}$ map (Fig. 5a). A similar area is observed in the cognitive paradigm's SPM$\{F\}$ (−38, −40, 6, F = 7.40; Fig. 6b); again, this area is not present on the fixed-effects SPM$\{Z\}$ (Fig. 6a). There was some overlap between voxels which displayed significant variability in each SPM$\{F\}$ and the corresponding fixed-effects SPM$\{Z\}$: for example, posterior SMA (−2, −8, 52; F = 4.67), ipsilateral cerebellum (26, −38, −22; F = 5.68), and contralateral precentral gyrus (−26, −18, 70; F = 4.66). These voxels were typically located at the edge of a larger cluster of activated voxels. The variability seen may reflect subtle differences in the areal extent of activations at the periphery of large clusters.

Fixed-Effects Analyses

The fixed-effects analyses of all three tasks (Figs. 5a, 6a, and 7a) displayed areas of activation concordant with previous studies employing a similar task. A number of fMRI studies have used finger-tapping as a stereotypical motor task (e.g., Rao *et al.*, 1993), and we found similar results (Table 1a), including the contralateral SM1 (Talairach coord. −38, −10, 52; Z score = 9.77), the anterior lobe of the ipsilateral cerebellum (20, −54, −18; Z = 9.50), the SMA (−2, −2, 52; Z = 9.39), the contralateral thalamus (−12, −18, 2; Z = 8.84), and the ipsilateral premotor cortex (38, −8, 50; Z = 8.80). It is notable that the SPM{*Z*} also contains areas not previously reported as activated by a simple externally paced finger-tapping paradigm, such as the right inferior parietal lobule (50, −28, 24; Z = 8.54). This is not surprising, as a single α_i of sufficient magnitude may be adequate for any voxel to pass the average significance threshold over sessions and so appear on the multisession fixed-effects SPM{*Z*}. If the fixed-effects SPM{*Z*} is viewed in isolation, it is impossible to know if these are "true" activated voxels which have not been reported in previous studies due to a lack of sensitivity or areas which display an activation effect significantly large to appear in the multisession fixed-effects maps, yet are not consistently activated across sessions.

Similar patterns of results were observed upon inspection of the multisession fixed-effects SPM{*Z*}'s from the cognitive and visual paradigms (Figs. 6a and 7a). Although less is known about the functional neuroanatomy of paced random-number generation, we found areas similar to a previous study (Table 2a; Jahanshahi *et al.*, submitted for publication). In common with Jahanshahi and colleagues, we found activation in the anterior cingulate cortex, but again noticed discrepancies between our results and theirs, e.g., our finding of bilateral calcarine cortex (12, −78, 8, Z = 6.45 and −6, −72, 20, Z = 5.80) and left SMA activation (−2, 18, 50, Z = 8.30). Similarly, our visual paradigm activated, as expected, striate and extrastriate areas around the calacarine sulcus (Fig. 7a, Table 3a), including bilateral V1 (14, −86, 2, Z = 9.72 and −8, −82, 0, Z = 9.57), in common with studies employing a comparable stimulus (e.g., Kwong *et al.*, 1992). However, as with the other paradigms, a number of areas not previously implicated in the functional neuroanatomy of this task were activated (e.g., the right SMA: 2, 2, 64, *Z* = 5.50).

Clearly, these effects beg closer scrutiny. If we wish to examine repeated trials of the same activation paradigm within a particular subject, it is necessary to define *variability* within the same subject. The fixed-effects SPM{*Z*}'s tell us where voxels are active on average across the observed sessions. If we wish to examine the generality of a single fMRI session, the variability of each voxel across all sessions must be addressed.

Random-Effects Analyses

Figures 5c, 6c, and 7c show random-effects analyses of each multisession data set. These SPM{*Z*}'s have been weighted by both between-session and withinsession variances of each data set. Upon visual inspection, the random-effects SPM{*Z*}'s resemble a "cleaned-up" version of the fixed-effects SPM{*Z*}'s, and each paradigm's pattern of results is now more in concordance with previous studies. There are still, however, areas within the random-effects SPM{*Z*} that one would not expect, *a priori*, to be involved in the functional neuroanatomy of each task (Tables 1b, 2b, and 3b). For example, the motor random-effects SPM{*Z*} (Fig. 5c) displays prominent bilateral auditory cortex activation (−42, −28, 18, Z = 8.14 and 48, −18, 14, Z = 6.16). We did not expect this, as pacing tones were played during both rest and activation epochs during each motor session. This result may reflect attentional modulation of auditory areas (Woodruff *et al.*, 1996; Grady *et al.*, 1997), as the tones' salience was different between the rest and the activation conditions. The neurobiological explanation for this result need not concern us here: it is sufficient to recognize that we did not predict this pattern of activation. If we had access to only a single session from our subject, we would have been suspicious about their true nature. Even a multisession fixed effects would not have helped: we would not be able to identify if the activation was driven by a small number of sessions only or was indeed a true positive. This reasoning demonstrates that multiple scanning sessions analyzed with an appropriate statistical model can reduce ambiguous interpretations.

The majority of voxels present in both the fixed-effects SPM{*Z*} and SPM{*F*} do not appear in the random-effects SPM{*Z*}'s. Properly accounting for between-session variance means that these voxels no longer survive a threshold of $P < 0.05$, corrected for multiple comparisons. This demonstrates that combining multiple sampling of sessions with a statistical model with more than one component of variance correctly accounts for even small session-by-condition interactions.

Figures 8 and 9 show voxels that typify different patterns of behavior across sessions, using the motor paradigm as an example. Figures 8A and 9A show a voxel in posterior SMA (−2, −8, −52) which survives a threshold of $P < 0.05$, corrected for multiple comparisons, in our multisession fixed-effects analysis. However, this voxel displays significant session-by-condition interactions (as seen by its appearance on the ESS-*F* map) and thus fails to survive correction when a

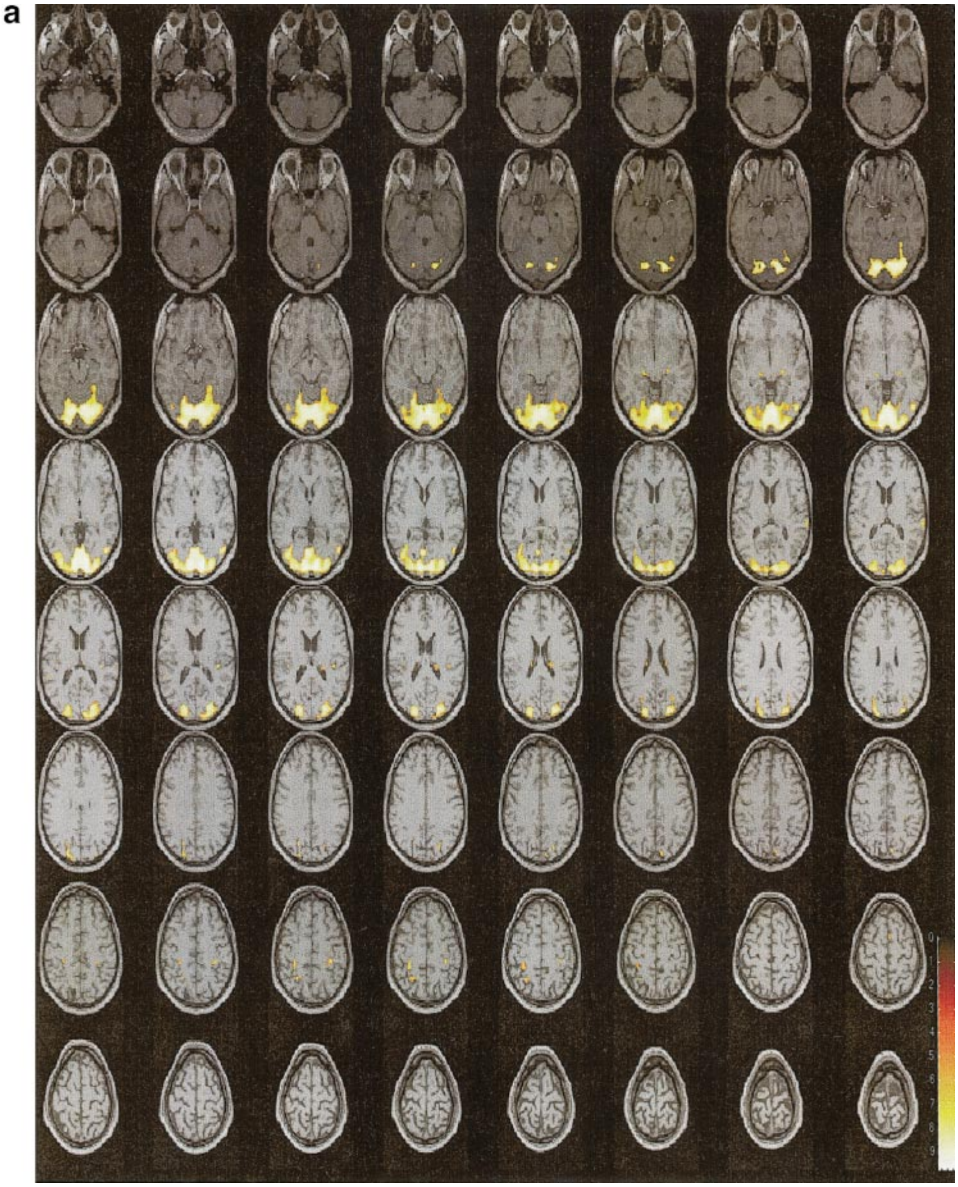

FIG. 7. Multisession analyses of the visual paradigm, analyzed using a fixed-effects model (a), extra-sum of squares *F* test (b), and a random-effects model (c). Voxels surviving the statistical threshold are displayed on a coregistered structural scan of the subject to aid the identification of activated areas. Each transverse slice is 2 mm thick. The color bar represents statistical significance, with higher *Z* and *F* scores having a brighter color.

random-effects model is used. This voxel is an excellent example of variability in "active" voxels. When one examines its parameter estimates by session (Fig. 9A), it is striking how stable it appears over some sessions (for example, sessions 15 to 18) and yet how variable its behavior is over all sessions. The histogram of parameter estimates in Fig. 9A shows that although only one session has a parameter estimate of greater than 1.5, this can still weigh the average activation effect over all sessions. When the variability of responses over sessions is addressed in the random-effects analysis, the voxel loses significance.

The voxel in left primary motor cortex (-36, -10, 52) displayed in Figs. 8B and 9B typifies voxels that survive statistical thresholds in both fixed- and random-effects analyses. This voxel shows remarkably similar parameter estimates over all sessions (Fig. 9B). The voxel in Figs. 8C and 9C is one that, although not significantly variable (not shown on the ESS$\{F\}$ map in Fig. 8C), does not survive correction when a random-effects model is used.

Voxels within each SPM$\{F\}$ can be thought of as belonging to various classes: those which are not activated by each paradigm, but display high vari-

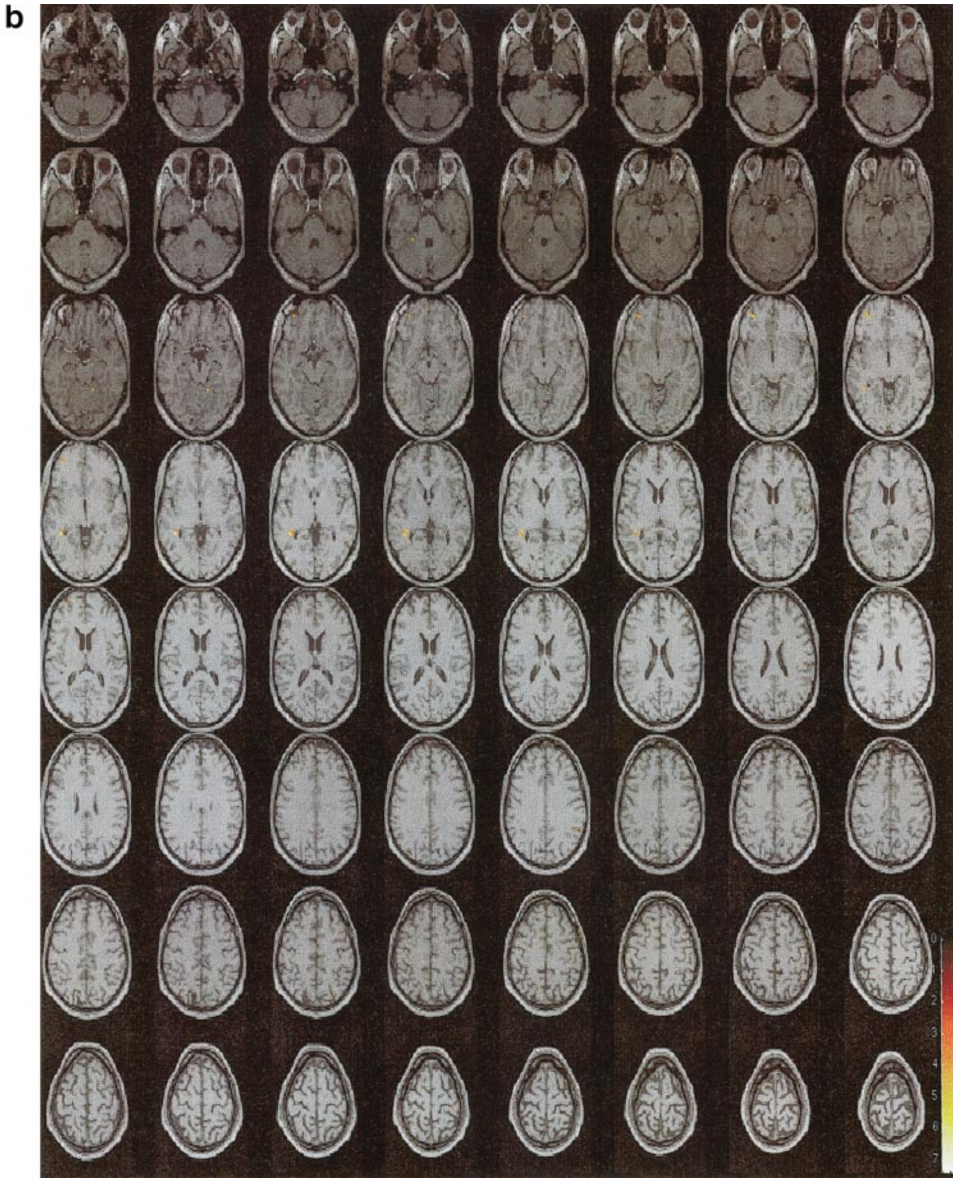

FIG. 7—*Continued*

ability of their parameter estimates (Figs. 8D and 9D); "true" active or deactivated voxels, surviving both fixed- and random-effects definitions of variability (Figs. 8B and 9B); voxels which are significant at a fixed-effects level but are significantly variable and do not survive correction for between-session variance (Figs. 8A and 9A); and voxels which, while not surviving a random-effects analysis, are not significantly variable as defined by the ESS{F} map (Figs. 8C and 9C).

DISCUSSION

The generality of any experimental result is an issue which confronts all researchers, independent of experimental discipline (Abelson, 1995). The results of *any* isolated experiment are always open to contamination, and fMRI is no exception. As fMRI is an ideal experimental technique to examine questions that require serial scanning sessions, there have been a number of previous studies that sought to examine the reproducibility of fMRI data. Researchers have examined similar activation paradigms across laboratories (Casey *et al.*, 1998), imaging modalities (Ojemann *et al.*, 1998),and sessions (Le *et al.*, 1997; Noll *et al.*, 1997; Rombouts *et al.*, 1998; Cohen *et al.*, 1999). These studies sought to characterize the reproducibility of fMRI data and so tried to ensure that each session was carried out similarly to those preceding it.

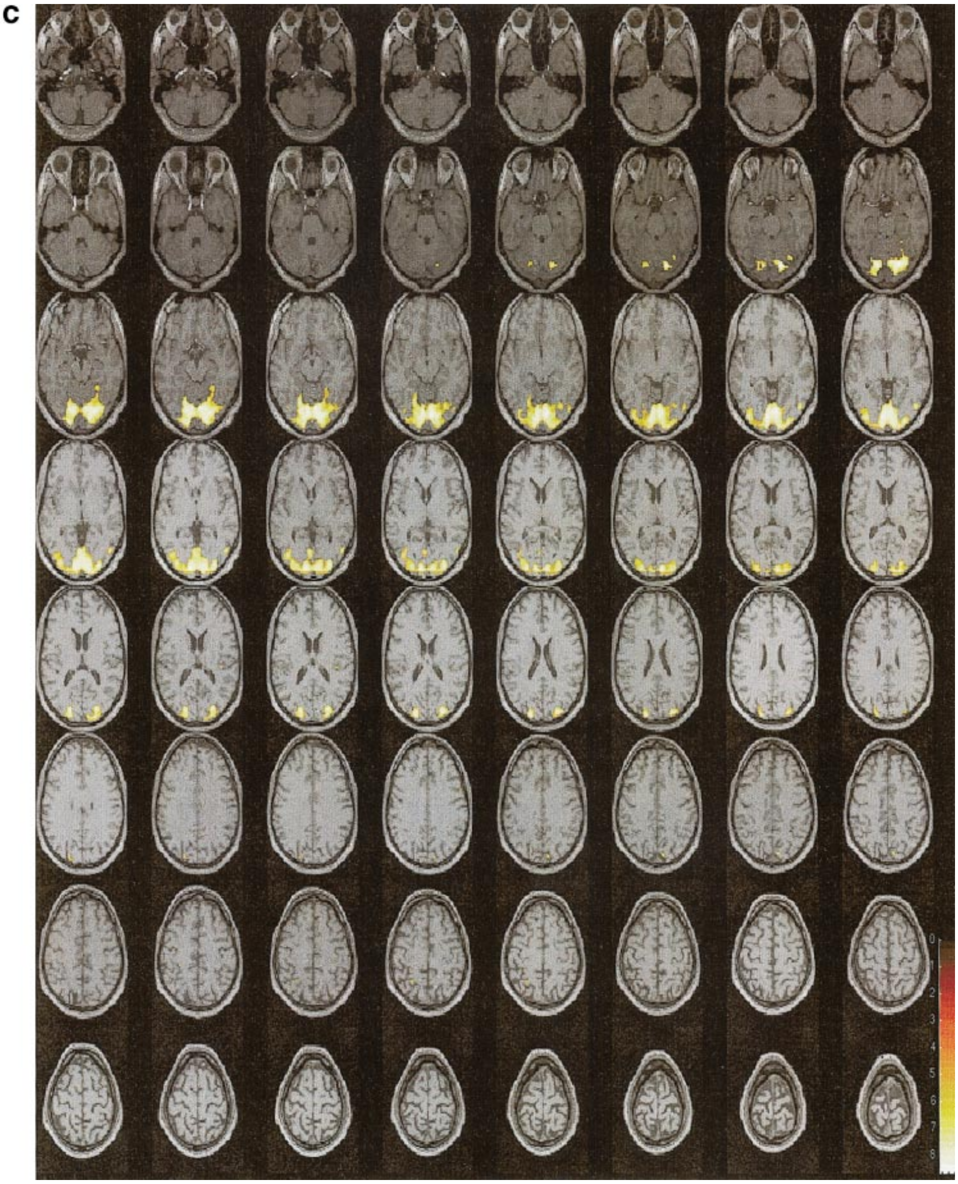

FIG. 7—*Continued*

Examining the reproducibility of fMRI data is an important question, but our own question was subtly different: we chose to examine how well a single-session data set from a single subject typified the subject's response across multiple sessions, using a variety of activation paradigms. By examining the variability in the magnitude of activation effects across a large number of sessions we accepted that each session would be different. Indeed, it was exactly this between-session variability that we wished to quantify.

We found that significant session-by-condition interactions occurred in each of the multisession data sets that we examined, as illustrated by the respective ESS SPM$\{F\}$'s. Our results are evidence of the influence of session context on the results of any individual session and show the potential danger of drawing general conclusions from an individual session analyzed in isolation with nothing known about reproducibility. If one samples more sessions, each successive session acquired facilitates a better estimation of between-session variance, thereby increasing power to detect the underlying response.

Differences in the Generality of Different Activation Paradigms

We chose to examine different activation paradigms to ensure that the results of our study would not be limited to a single class of activation task. The majority

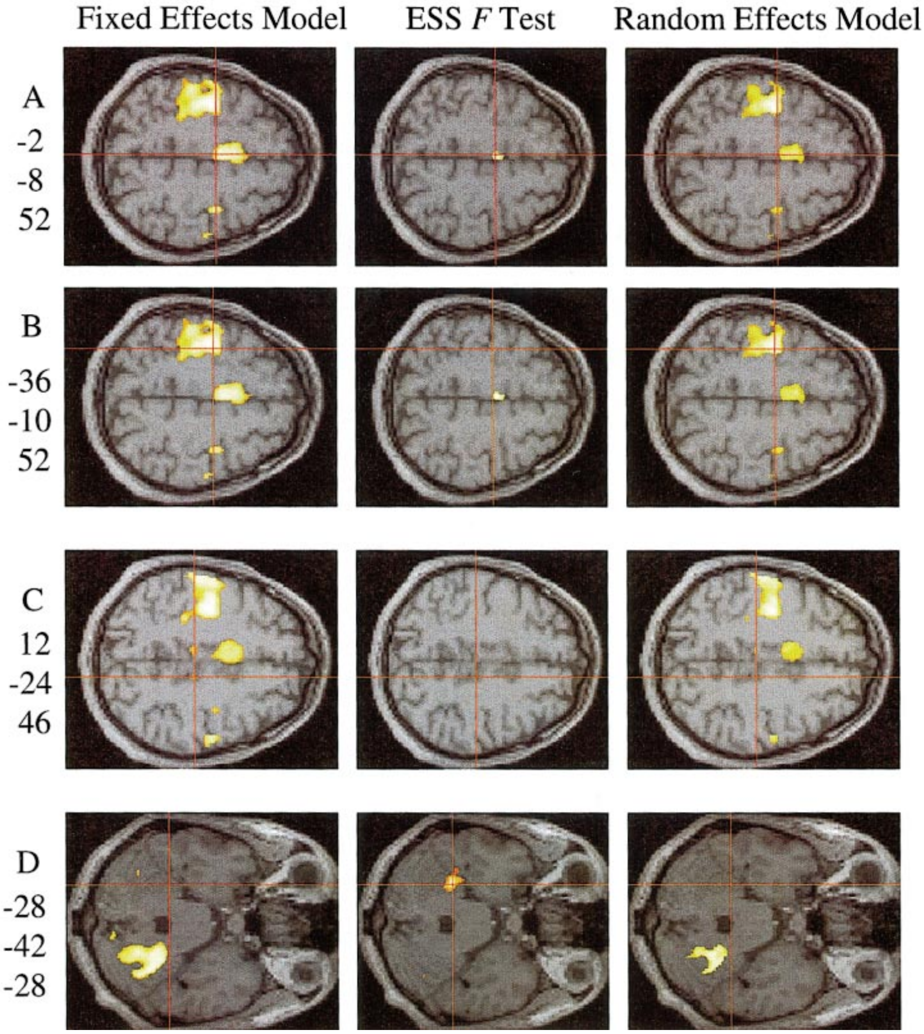

FIG. 8. Examples of voxels from the multisession motor analyses that typify different kinds of statistical significance. Voxel A is significant at a fixed-effects level, but variable enough to not appear in a random-effects analysis. Voxel B is significant at both a fixed- and a random-effects level and does not display significant session-by-condition interaction terms (it does not appear in the ESS-*F*). Voxel C is significant at a fixed-effects level and is not significantly variable across sessions, but does not possess a random-effects level of significance. Voxel D has significance at neither fixed- nor random-effects level, but is significantly variable between sessions to appear in the ESS-*F*.

of work examining repeatability in fMRI has employed simple visual or motor paradigms, though there has been limited use of "higher" cognitive activation paradigms (Yetkin *et al.*, 1996; Noll *et al.*, 1997; Casey *et al.*, 1998). Our initial expectations were that the visual task would prove to have the fewest session-by-condition interactions and the cognitive task the most. While we did not compare the number or magnitude of voxels in each paradigm's SPM{*F*}, we were initially surprised to note that the visual SPM{*F*} appeared to have prominent bilateral areas of high variability in primary visual cortex (Fig. 7b), while the cognitive SPM{*F*} contained few voxels which overlapped with areas activated consistently by the activation task itself. As mentioned above, the visual activations were located in areas not activated on average by the task itself, but it is still surprising to find them in such close proximity to primary visual cortex. We suggest that slight differences in visual field coverage by our visual stimulus may have caused these effects, producing a variable rim around a core of visual cortex that was consistently stimulated across sessions. A further possibility is that these results reflect the high concentration of venules in the microvasculature of visual cortex (Marinković *et al.*, 1995), which may cause a higher variability in its response to afferent stimulation.

Sources of Session-by-Condition Interactions

We did not attempt to systematically assess the relative magnitudes of different sources of variance on session reproducibility, as in some previous studies (e.g., Noll *et al.*, 1997). We acknowledge, however, that identifying the sources of intersession variance

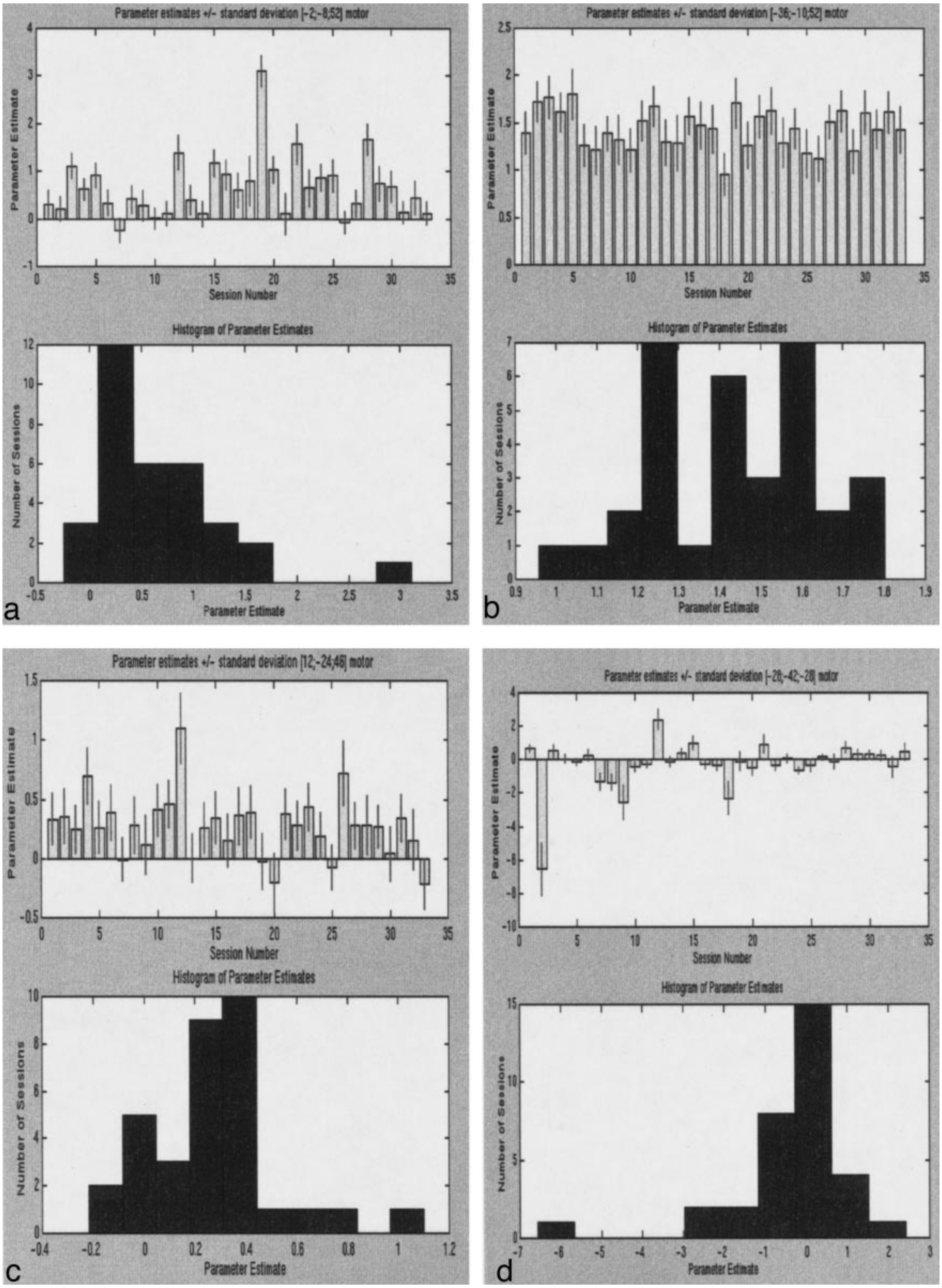

FIG. 9. Session-by-session plots of the parameter estimates (α_i) and their standard deviations (vertical bars) of the voxels from Fig. 8. The histograms below each plot show the spread of values of α across all sessions.

is important. For example, it is possible that spatial preprocessing may affect intersession variance quite independent of underlying physical or physiological variability. The realignment procedure used (Ashburner *et al.*, 1997) seeks to minimize the sum-of-squares differences between successive volumes and a reference (here, the first volume in the time series). It is a facile point that each paradigm induces intensity changes in voxels (i.e., it "activates" them), and so volumes acquired during the "on" period of each paradigm will contain focal intensity differences from volumes acquired during "rest." As a successful realignment between two volumes relies on the volumes used being similar rigid bodies, differing only in their alignment in space, the paradigm-induced intensity changes will affect the efficacy of alignment and may ultimately raise intersession variance. In addition, similar effects in voxels lying at a tissue boundary (voxels in the walls of the ventricles, for example) may result from simple repositioning of the subject between sessions, causing session-specific partial volume effects.

Examples like the above make it difficult to conclude if the variability which we observe is attributable to differences in: (i) the scanning environment (e.g., position of subject within headcoil) or (ii) preprocessing (misalignment). We are more confident that we were able to minimize session-by-condition interactions that could be attributable to performance differences. We treated each successive scanning session as though it was the first time that our subject had been scanned, to examine the potential influence of session context on a single-session experiment (acknowledging that we cannot overcome the Groundhog Day effect). It could be argued that systematic differences in our subject's performance across sessions may have resulted in the session-by-condition interactions which we observed, as the repeated execution of any active task or protocol of sensory stimulation may result in habituation or learning effects (e.g., Karni *et al.*, 1995). We chose our activation tasks to minimize this possibility. Our subject was pretrained on the motor task, and our task frequency was chosen to lie within a range previously demonstrated by Blinkenberg and colleagues (1996) to have a low error rate (between 1 and 2 Hz). Similarly, we chose a stable rate of number generation for our cognitive task (informed by the results of Jahanshahi *et al.*, submitted for publication). Although we did not record subject performance on this task (primarily because of the motion it would produce), performance at random number generation remains stable over a number of repetitions (Evans *et al.*, 1980). Furthermore, a preliminary MANCOVA analysis of our motor data set examining the effects of session revealed no systematic expression of the experimental variance over subsequent sessions (data not shown). Although we accept that learning effects may exhibit complex temporal dynamics, the structure of our parameter estimates over sessions suggests random variation (Fig. 8) around a "true" mean parameter estimate. However, without independent measures of task performance, we cannot entirely rule out between-session habituation or learning-related changes in activation.

Stability of fMRI Results across Sessions—Consequences for Longitudinal Studies

As noted previously, fMRI is ideally suited to the examination of learning or recovery-of-function studies. These studies are typically predicated on the assumption that the experimental effects will be large enough to ensure their detection compared to nonspecific between-session effects. Because of the considerable time that must typically be devoted to such studies, it would be useful to have some idea of the relative magnitudes of each effect before beginning. Although in the present study we have examined the similarity of the results between sessions while accepting a certain degree of difference in the practical implementation of each session, our results do not really address the issue of signal to noise in longitudinal fMRI studies. As *any* difference between sessions is a session-by-condition interaction, any study which purports to focus on session-by-condition interactions produced by the experimental manipulation must ensure that nonspecific session-by-condition interactions can be efficiently controlled (for a discussion of these issues, see Petersson *et al.*, 1999). We examined three tasks that were designed to show limited session-by-condition interactions in subject performance. As such, our results cannot be used to address the validity of longitudinal fMRI studies. We note, however, that the stability of our results suggests that longitudinal studies that produce unambiguous results should be feasible.

Use of Thresholded Statistical Maps to Analyze Session Generality

Typically, the results of neuroimaging experiments are displayed using binarized statistical maps. In this fashion, voxels that pass a predetermined statistical threshold are classified as active and other voxels as inactive. Although the utility and clarity of the results motivate this approach, much of the richness of functional neuroimaging data sets is removed. Attempts to examine the test–retest reliability of fMRI using measures such as "voxel counting" on thresholded maps therefore suffer from two problems: an essentially arbitrarily defined statistical threshold and the loss of complexity which accompanies any method that has to classify voxels as either active or inactive. We sought, instead, to characterize our data sets in terms of the between-session variance of the activation effects and not merely examine when voxels passed an arbitrarily set threshold on successive sessions. The differences between the two approaches are apparent when one compares the results of our single-session analyses (Figs. 2, 3, and 4) with our later multisession analyses (Figs. 5, 6, and 7). Generally a failure to detect activation may say more about the sensitivity of the experiment than the presence of the effect itself (Poline *et al.*, 1996). Certain areas may therefore appear more variable than they truly are.

Effects of Sample Size on the Analysis of Generality

Our results demonstrate the need for a large sample size when examining how well a single fMRI session exemplifies a subject's responses. The plots of parameter estimates by session in Fig. 9 show that voxels in which we found significant session-by-condition interactions over all of our sessions appear surprisingly stable when examined over a small number of sessions (for example, sessions 27–29 in Fig. 9C are almost identical). This effect has been termed "the law of small numbers" (Tversky *et al.*, 1971)—the tendency to ascribe a lack of variability to small sample groups. Our use of a large number of sessions allowed us to characterize variability that may have been missed by previous studies employing five repeated sessions at most on the same subject. However, the opposite argument may be leveled at our results: if the sample size is large enough, then a statistically significant difference will always be found—this is merely an example of the fallacy of classical hypothesis testing. We accept this criticism, but believe that an analysis of 30 sessions is an appropriate sample size for the purposes of this study. The existence of significantly variable voxels necessitates the use of a random-effects model to allow the experiments to truly generalize their results to the subject.

Levels of Inference Arising from Fixed- and Random-Effects Models

Worlsey and colleagues (1992) first suggested the use of a "summary statistic" approach to the analysis of functional neuroimaging data. However, the implementation we used in the current study is that of Holmes and Friston (1998), who suggested random-effects analyses for balanced designs in neuroimaging employing a general linear framework to allow for the between-subject variance component in multisubject designs. As discussed previously, the random-effects analysis confers generality, but with a concomitant loss of sensitivity due to the inevitable low degrees of freedom. We assumed that the ϵ_i were Normally distributed and incorporated this assumption into our random-effects level model. However, by examining Fig. 9 it is clear that the ϵ_i do not necessarily conform to this distribution. If we examine the case of the voxel in Fig. 9A, it is clear that this voxel has a skewed-right distribution. Indeed, if one asks a simpler question of the voxel in Fig. 9A (how often is $\alpha > 0$) and employ a simple sign test, the probability of getting 31/33 positive α's is $<7 \times 10^{-8}$. Yet this voxel does not pass the random-effects analysis used here. Although we accept that this is only one voxel, it casts doubts on assumptions of Normality for the ϵ_I, and it is clear that further investigation is needed into the distribution of between-session variance. The development of random-effects models that do not require prior assumptions of the distribution of residuals may be needed to address this issue.

The use of random-effects models in the analysis of fMRI data is a recent addition to the canon of neuroimaging analysis methods, and it is wise to note a previous adoption of this measure. In the analysis of behavioral data from human subjects, Clark's (1973) initial proposal that the model should be used more frequently highlighted an obvious problem: *treating* a sample from a population as random and *selecting* a sample randomly from a population are clearly not the same. Although we could argue that by using a random-effects analysis in our study, we can generalize our results to our subject as a putative population, we have performed a very limited sampling of our subject's responses. Each scanning session was performed over a 2-month period only, and session times were selected in a biased manner: near midday and near 6 PM in the evening. However, it is not elegant to have to state that "our results generalize to the population of possible sessions sampled from our subject over a period of 2 months, using the resources available in our laboratory." In practice, these caveats are usually accepted. Indeed, the use of random-effects models to ensure the correct level of inference in multisubject fMRI analyses rarely addresses the other sources of systematic variation in the population that the investigators are generalizing to (usually male, Caucasian right-handers who respond to advertisements and financial reward). However, adopting a random-effects model does afford some protection against inappropriate generalization of results, as noted by Abelson (1995).

Although we have shown that with an appropriate statistical model and a large sample of sessions we can obtain robust results, a number of issues remain unanswered. In particular, we would hesitate before generalizing our own results to other centers, subjects, or activation paradigms, as between-session variance may vary greatly depending on the context under which it is studied.

CONCLUSIONS

In this paper we described the results of an experiment designed to examine intersession variance in fMRI during the performance of simple visual, motor, and cognitive tasks by a single subject. First, analyzing our data session by session, we suggested that binarized statistical maps, though convenient, are not a useful tool for the evaluation of intersession variability. We then described an analytical framework that allowed us to identify significantly variable voxels by session across our multisession data sets. Each multisession data set, by paradigm, showed evidence of significant session-by-condition interactions. This result demonstrates that session context effects have a significant effect on fMRI data and illustrates that a single session should be considered merely as a single sample of a subject's responses to the experimental intervention employed. As we sampled a large number of sessions across all paradigms, we then compared the differences between analyzing these data using either fixed- or random-effects linear models, the latter being a recent addition to neuroimaging analysis. Although we comment on the usefulness of random effects analyses, which allow inference about experimental effects to be extended to the population which the sessions were sampled from, we draw attention to the intersession distribution of voxel response amplitudes. Our assumption of Normally distributed intersession residuals was not supported by close examination of some of our data, and so we accept that future work is required before random-effects models can be used to their full potential. Finally, we acknowledge that identifying the source and magnitude of the different sources of intersession variance in fMRI is crucial. The ability to differentiate between variability caused by the neurovascular signals that fMRI measures, and variability introduced by the means of measurement and analysis of these signals, is essential for the future of fMRI as a noninvasive imaging modality.

ACKNOWLEDGMENTS

We thank Jesper Andersson and John Ashburner for invaluable help and advice and our reviewers for their constructive and generous comments. The Wellcome Trust supported this work.

REFERENCES

Abelson, R. P. 1995. *Statistics as Principled Argument.* Elbaum, Hillsdale, NJ.

Ashburner, J., and Friston, K. J. 1999. Nonlinear spatial normalisation using basis functions. *Hum. Brain Mapp.* **7:**254–266.

Ashburner, J., Neelin, P., Collins, D. L., Evans, A. C., and Friston, K. J. 1997. Incorporating prior knowledge into image registration. *NeuroImage* **6:**344–352.

Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R. M., Cohen, M. S., Vevea, J. M., Brady, T. J., and Rosen, B. R. 1991. Functional mapping of the human visual-cortex by magnetic-resonance-imaging. *Science* **254:**716–719.

Blinkenberg, M., Bonde, C., Holm, S., Svarer, C., Andersen, J., Paulson, O. B., and Law, I. 1996. Rate dependence of regional cerebral activation during performance of a repetitive motor task: A PET study. *J. Cereb. Blood Flow Metab.* **16:**794–803.

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

Clark, H. H. 1973. The language-as-fixed-effect fallacy. *J. Verb. Learn. Verb. Behav.* **12:**335–359.

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

Draper, N. R., and Smith, H. 1981. *Applied Regression Analysis,* pp. 97–98. Wiley, New York.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., and Peters, T. M. 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In *Proceedings, IEEE Nuclear Science Symposium and Medical Imaging Conference,* pp. 1813–1817. IEEE Inc., Piscataway, NJ.

Evans, F. J., and Graham, C. 1980. Subjective random number generation and attention deployment during acquisition and over-learning of a motor skill. *Bull. Psychonom. Soc.* **15:**391–394.

Fox, P. T., and Raichle, M. E. 1985. Stimulus rate determines regional blood flow in striate cortex demonstrated by positron emission tomography. *Ann. Neurol.* **17:**303–305.

Frison, L., and Pocock, S. J. 1992. Repeated measures in clinical trials: An analysis using mean summary statistics and its implications for design. *Stat. Med.* **11:**1685–1704.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., and Frackowiak, R. S. J. 1995. Spatial registration and normalisation of images. *Hum. Brain Mapp.* **2:**165–189.

Genovese, C. R., Noll, D. C., and Eddy, W. F. 1997. Estimating test–retest reliability in functional MR imaging. I. Statistical methodology. *Magn. Res. Med.* **38:**497–507.

Grady, C. L., Van Meter, J. W., Maisog, J. M., Pietrini, P., Krasuski, J., and Rauschecker, J. P. 1997. Attention-related modulation of activity in primary and secondary auditory cortex. *NeuroReport* **28:**2511–2516.

Henson, R. N. A., Shallice, T., and Dolan, R. J. 1999a. Right prefrontal cortex and episodic memory retrieval: A functional MRI test of the monitoring hypothesis. *Brain* **122:**1367–1381.

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., and Dolan, R. J. 1999b. Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *J. Neurosci.* **19:**3962–3972.

Holmes, A. P., Josephs, O., Büchel, C., and Friston, K. J. 1997. Statistical modeling of low frequency confounds in fMRI. *NeuroImage* **5:**S480.

Holmes, A. P., and Friston, K. J. 1998. Generalisability, random effects and population inference. *NeuroImage* **7:**S754.

Howseman, A. M., McGonigle, D. J., Grootoonk, S., Ramdeen, J., Athwal, B. S., and Turner, R. 1998. Assessment of the variability in fMRI data sets due to subject positioning and calibration of the MRI scanner. *NeuroImage* **7:**S599.

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

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner, R., Cheng, H. M., Brady, T. J., and Rosen, B. R. 1992. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA* **89:**5675–5679.

Le, T. H., and Hu, X. 1997. Methods for assessing accuracy and reliability in functional MRI. *NMR Biomed.* **10:**160–164.

Marinković, R., Cvejin, B., Marković, L., and Budimlija, Z. 1995. [Morphologic characteristics of the vascular network in the striate area in humans]. *Med. Pregl.* **48:**7–9.

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

Ogawa, S., Menon, R. S., Tank, D. W., Kim, S. G., Merkle, H., Ellerman, J. M., and Urgubil, K. 1992. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* **89:**5951–5955.

Ogawa, S., Menon, R. S., Kim, S.-G., and Ugurbil, K. 1998. On the characteristics of functional magnetic resonance imaging of the brain. *Annu. Rev. Biophys. Biomol. Struct.* **27:**447–474.

Ojemmaan, 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., and 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.* **6:**203–215.

Petersson, K. M., Elfgren, C., and Ingvar, M. 1999. Learning-related effects and functional neuroimaging. *Hum. Brain Mapp.* **7:**234–243.

Poline, J. B., Vandenberghe, R., Holmes, A. P., Friston, K. J., and Frackowiak, R. S. J. 1996. Reproducibility of PET activation studies: Lessons from a multi-center European experiment. *NeuroImage* **4:**34–54.

Rao, S. M., Binder, J. R., Bandettini, P. A., Hammeke, T. A., Yetkin, F. Z., Jesmanowicz, A., Lisk, L. M., Morris, G. L., Mueller, W. M., Estkowski, L. D., Wong, E. C., Haughton, V. M., and Hyde, J. S. 1993. Functional magnetic resonance imaging of complex human movements. *Neurology* **43:**2311–2318.

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

Searle, S. R., Casella, G., and McCulloch, C. E. 1992. *Variance Components.* Wiley, New York.

Talairach, P., and Tournoux, J. 1988. *A Stereotactic Coplanar Atlas of the Human Brain.* Thieme, Stuttgart.

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

Tversky, A., and Kahneman, D. 1971. Belief in the ‘law of small numbers.’ *Psychol. Bull.* **75:**105–110.

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

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12:**900–918.

Worsley, K. J., and Friston, K. J. 1995. Analysis of fMRI timeseries revisited—Again. *NeuroImage* **2:**173–181.

Yetkin, F. Z., McAuliffe, T. L., Cox, R., and Haughton, V. M. 1996. Test–retest precision of functional MR in sensory and motor task activation. *AJNR* **17:**195–198.