Published in final edited form as:
Neuroimage. 2010 October 15; 53(1): . doi:10.1016/j.neuroimage.2010.02.084.

Multi-site characterization of an fMRI working memory paradigm: Reliability of activation indices

Anastasia Yendiki[1,*], **Douglas N. Greve**[1], **Stuart Wallace**[1,2], **Mark Vangel**[1,3], **Jeremy Bockholt**[4,5], **Bryon A. Mueller**[6], **Vince Magnotta**[7], **Nancy Andreasen**[7], **Dara S. Manoach**[1,2], and **Randy L. Gollub**[1,2]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Dept. of Radiology, MGH, Dept. of Radiology, Harvard Medical School

[2]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

[3]Mallinkcrodt GCRC Biomedical Imaging Core, Massachusetts General Hospital, Charlestown, MA, USA

[4]University of New Mexico, Albuquerque, NM, USA

[5]The Mind Research Network, Albuquerque, NM, USA

[6]Department of Psychiatry, University of Minnesota, Minneapolis, MN, USA

[7]University of Iowa, Iowa City, IA, USA

Abstract

Neuroimaging studies are facilitated significantly when it is possible to recruit subjects and acquire data at multiple sites. However, the use of different scanners and acquisition protocols is a potential source of variability in multi-site data. In this work we present a multi-site study of the reliability of fMRI activation indices, where 10 healthy volunteers were scanned at 4 different sites while performing a working memory paradigm. Our results indicate that, even with different scanner manufacturers and field strengths, activation variability due to site differences is small compared to variability due to subject differences in this cognitive task, provided we choose an appropriate activation measure.

Keywords

fMRI; multi-center studies; reliability; SIRP

1. Introduction

Multi-site studies provide an efficient means for collecting neuroimaging data from a large number of subjects. Thus they augment our ability to study conditions that are relatively rare in the general population, allow larger samples for studies of genetic polymorphisms, and increase the generalizability of the findings. However, differences in scanner hardware and

[*]Author to whom correspondence should be addressed: Anastasia Yendiki, Athinoula A. Martinos Center for Biomedical Imaging, 149 13th St., Charlestown, MA 02129, tel: 617 726-9434, fax: 617 726-7422, ayendiki@nmr.mgh.harvard.edu.

acquisition protocols may be a source of variability in the data. It is important to quantify this effect and compare it to the variability introduced by other factors, such as individual subject differences and imaging noise, before embarking on studies where data are pooled across multiple sites.

Several recent studies have shown fMRI activation measures to be highly reproducible across sites with identical scanners in tasks ranging from facial affect processing [Suckling 2008] to motor [Costafreda 2007, Sutton 2008] and visual [Sutton 2008. In particular, these studies have found the proportion of the variance in activation measures that can be attributed to across-site variability to be an order of magnitude smaller than the proportion that can be attributed to across-subject variability.

Pooling data acquired at sites with different scanners poses additional challenges. Initial results from a multi-site study performed by the Biomedical Informatics Research Network (BIRN, http://www.nbirn.net) indicated that scanner differences could result in significant variability in fMRI-derived measures of brain activation [Zou 2005]. These results were obtained for a sensorimotor paradigm, performed by 5 subjects at 10 different scanners. The experience from this study led to a series of recommendations on how to mitigate across-site variability. These include a quality assurance protocol to ensure stable scanner performance [Friedman 2006a] and guidelines for data analysis methods that lead to improved reliability of activation measures [Friedman 2008].

In this work we present results from a study of neuroimaging data reliability conducted by the Mind Research Network (MRN) sponsored Mind Clinical Imaging Consortium (MCIC). For this study 10 healthy volunteers traveled to 4 sites and were scanned twice. Structural, functional, and diffusion-weighted MRI data were acquired at each site. Here we focus on the reliability of the functional data.

At the time of the study the sites had scanners from different manufacturers (GE, Waukesha, WI, USA or Siemens, Erlangen, Germany) and with different field strengths (1.5T or 3T). However, all of the sites are also members of BIRN and thus the present study benefited from the lessons learned by phase I of the BIRN study in addressing some of the factors that may result in site differences. This effort included following the specifications of the quality assurance protocol proposed by the BIRN [Friedman 2006a], as well as standardizing certain acquisition parameters across sites, as described in more detail later.

Although the study presented here involved healthy subjects, it was performed with the ultimate goal of informing a large-scale, multi-site fMRI study of schizophrenia conducted at the same four sites by the MCIC. To this end, the paradigm studied here is one of particular interest to schizophrenia research. It consisted of a variation of the Sternberg item recognition paradigm (SIRP) [Sternberg 1966], tailored for use in neuroimaging experiments [Manoach 1997]. Performance of the SIRP is relatively stable in healthy participants, even after extensive daily practice [Kristofferson 1972]. In fMRI studies, the SIRP gives rise to activation in a network of brain areas associated with working memory and has been used to characterize working memory deficits in schizophrenia patients [Manoach 1999, Manoach 2000, Ragland 2007]. The within-subject reliability of SIRP activations has been found to be high for healthy subjects but low for schizophrenia patients [Manoach 2001]. Here we study the across-site reliability of these activations in healthy individuals.

2. Materials and methods

I. Experimental design and data acquisition

Ten healthy subjects (ages 30-63, 5 male) traveled to four sites and were scanned while performing the SIRP on each of two visits (test-retest). The four sites were: Massachusetts General Hospital (MGH), University of New Mexico (UNM), University of Iowa, and University of Minnesota. Two of the sites used 3T scanners (Siemens at MGH and Minnesota), while the other two used 1.5T scanners (Siemens at UNM and GE at Iowa).

The participating sites are also members of the BIRN and in that capacity they had been part of multi-site MRI calibration studies by the Morphometry BIRN [Han 2006, Jovicich 2009] and Function BIRN [Freidman 2006, Freidman 2008]. The lessons learned from those studies were then applied to reduce disparities in the experimental set-up, data acquisition methods and sequences used for the study presented here. In particular, all sites had matched button press devices, followed common audiovisual set-up calibration methods and paid particular attention to centering each subject's head in the center of the scanner bore to minimize gradient distortion effects. Sequences parameters such at bandwidth and echo spacing were optimized at each site for the best quality images and synchronization of the stimulus onset with the scan start was improved. In addition, the four sites followed the quality assurance procedures recommended by the BIRN to ensure scanner stability [Friedman 2006a]. However, each site followed its own choice of head immobilization strategy (foam packing, soft-strap restraints, or none). The subjects wore Avotec headphones with active noise cancellation (Avotec, Inc., Stuart, FL) during all scans at all sites.

During each visit, a subject performed the SIRP task (EPrime v1.1, Psychology Software Tools, Inc., Pittsburg, PA) during four separate scans. Thus each subject performed the task paradigm a total of 32 times (4 scans × 2 visits × 4 sites). A total of 316 scans were analyzed because data was not available for one of the visits of one of the subjects. Most of the test-retest visits took place on subsequent days. The only exceptions were two cases with 2 days between test and retest and one case each with 6, 7, and 32 days between test and retest.

For each scan we acquired whole-brain, gradient-echo, EPI data along 27 contiguous oblique axial slices, parallel to the AC-PC line (in-plane resolution 3.44mm, slice thickness 4mm skip 1mm, slice order interleaved, TE=30msec for 3T, TE=40msec for 1.5T, TR=2sec, FA=90°, FOV=22cm). A total of 177 time frames were collected for a total scan time of 5min 54sec.

During each scan the subject had to retain in memory a set of 1, 3 or 5 digits during blocks of 46sec, providing a range of task difficulty. First the subject was prompted by the word “Learn” for a time of 1.5sec (*prompt* condition), followed by a blank screen for 0.5sec. Then the targets (digits to be retained in working memory) were presented in red font for a time of 6sec (*encode* condition). The subject was then shown a sequence of probe digits in green font and had to indicate whether each probe digit was a target or a foil, *i.e.*, whether it was a member of the memorized set or not (*probe* condition). The probe condition lasted a total time of 38sec. Each probe digit was presented for up to 1.1sec in a pseudo-randomly jittered fashion within a 2.7s interval. We presented 14 probe digits in each block, of which 7 were targets and 7 were foils, for a total of 84 probes per scan. Subjects were instructed to respond with a right-thumb button press if the probe digit was a target and a left-thumb button press if it was a foil.

Subjects were instructed to respond as quickly and as accurately as they could. They were told that they would receive a bonus of $0.05 for every correct response. Subjects were

trained to perform the task on a computer prior to the first scan session to verify that they achieved a greater than chance performance.

A working-memory (WM) block consisting of a single repetition of the *prompt-encode-probe* conditions was then repeated six times per scan. We alternated WM blocks with blocks of fixation. The durations of the fixation blocks were random integer multiples of 2sec, chosen so that the total duration of all fixation blocks within a scan was 78sec. Among the six WM blocks in a scan, there were two blocks of each of the three set sizes (1, 3, 5) in a pseudorandom order.

We varied the digits that comprised the memory sets for each of the 32 scans to eliminate learning effects. The target digits presented in each block were randomly chosen integers between 0 and 9, with no digit repeated within a single set. To avoid response biases, no digit was used more than 60% of the time as a target digit across the 6 scans in a visit (2 practice scans and 4 experimental scans). Also, in the two sets within a scan that consisted of a single digit, that digit was not the same. The order of targets and foils within a probe epoch was random, but no more than 3 consecutive digits could be targets. Each of the target digits presented during the encode epoch had to be presented at least once during the probe epoch. When the set presented during the encode epoch consisted of 3 target digits, each target digit had to be presented at least twice during the probe epoch.

In addition to the functional data, T1-weighted high-resolution structural scans were collected and we use them here for anatomical localization. Although T1-weighted scans were acquired at all four sites, we used the ones collected at a single site throughout the analyses presented here, as our focus in this work was the variability of the functional data. Specifically, the T1-weighted scans that we used in the present study were acquired at the MGH site on a Siemens 1.5T scanner with an axial GRE sequence (in-plane resolution 0.625 mm, slice thickness 1.5 mm, FOV=16 cm, 256×256×144 matrix, TR=12ms, TE=4.76ms, FA=20°, NEX=3).

II. Analysis of behavioral data

We recorded the accuracy and latency of the subjects' responses to the probe digits using the same equipment at all sites (EPrime and a NeuroScan response pad, NeuroScan, Charlotte, NC). We performed analysis of variance (ANOVA) on the reaction time (RT) data, modeling subject, site, and run as random effects, and visit, memory load (1, 3, 5), probe type (target or foil), and site order as fixed effects.

III. Analysis of fMRI data

(i) Quality assurance—We evaluated the quality of the fMRI data using the artifact detection tools (ART) [Whitfield-Gabrieli 2009]. The purpose of this evaluation was to ensure that no scans with gross motion or spiking artifacts were included in our data set, as this could have confounded our measures of across-site reliability. We checked for outlier time frames in each time series in our data set that satisfied any of the following criteria for exclusion: *(i)* Global mean image intensity that differed by more than 3 standard deviations from the mean of the entire series of time frames in a scan, *(ii)* Displacement due to motion by more than 1mm in the x, y or z direction relative to the previous time frame or *(iii)* Rotation due to motion by more than 0.1rad around any of the three axes relative to the previous time frame.

Out of the 316 total scans in the data set, 85 scans had no time frames flagged as outliers and only 10 scans had more than 10% of their time frames flagged as outliers. The average number of outliers per run was 1.6 ± 2.1 (UNM), 6.4 ± 11.4 (Iowa), 1.8 ± 2.6 (MGH), 1.9 ±

2.2 (Minn). Specifically, four of the ten subjects exhibited significantly more motion-related outlier time points at Iowa than at any of the other sites. This sort of variation between sites could be due to the differences in head immobilization strategies that may have led to different levels of subject comfort. In addition, subjects reported that late-night scans at Iowa were the hardest for concentration and comfort.

We repeated all analyses of variance discussed below after removing the outlier time frames through the use of nuisance regressors in the linear model and, in the case of scans where more than 10% of the time frames were flagged as outliers, removing the entire scan from the analysis. This had negligible impact on the outcome of our analyses, so we chose to include all time frames from all scans for the results reported in the following.

(ii) Statistical analysis—We used FEAT (FMRI Expert Analysis Tool), part of FSL (FMRIB's Software Library, http://www.fmrib.ox.ac.uk/fsl), to perform statistical analysis. We performed the following pre-processing steps on the acquired images: *(i)* Motion correction using MCFLIRT [Jenkinson 2002], *(ii)* Removal of non-brain voxels using BET [Smith 2002], *(iii)* Spatial smoothing using a 3-D Gaussian kernel with a FWHM of 5mm, *(iv)* Normalization of all volumes to a common average scan intensity, and *(v)* High-pass temporal filtering (Gaussian-weighted LSF straight-line fitting, with sigma=50.0s).

We analyzed each pre-processed time series using the Functional Imaging Linear Model (FILM) with local autocorrelation correction [Woolrich 2001]. We fit a general linear model to the series, including each of the *prompt, encode*, and *probe* conditions at each of the 3 memory loads as a separate explanatory variable. For all conditions, the haemodynamic response function was modeled as a single gamma function with initial delay 0sec, time-to-peak 6sec, and dispersion 3sec. Although motion parameter estimates could be included in the GLM as nuisance regressors [Friston 1996], we did not include them in our model. For the results shown here we used the following linear Contrasts Of Parameter Estimates (COPEs), where 1t, 3t and 5t signify blocks with memory sets of 1, 3, and 5 target digits respectively: *(i)* Probe-1t versus fixation, *(ii)* Probe-3t versus fixation, *(iii)* Probe-5t versus fixation, *(iv)* Probe-5t versus Probe-1t, and *(v)* Any load (Probe-1t or Probe-3t or Probe-5t) versus fixation. Each of the estimated COPEs and its estimated variance for each scan were used to obtain a T-statistic map.

To enable the use of ROIs defined anatomically on a subject-by-subject basis, we registered each subject's functional images to a high-resolution T1 image of the same subject collected at MGH. This was done by an intra-subject registration method that maximizes the intensity contrast gradient of the image across the cortical gray/white boundary, which is obtained from the T1 scan [Greve 2009]. To perform multi-subject analyses, we also registered the functional images to the standard space defined by the MNI-152 atlas [Talairach 1988]. We did this by first registering the T1 images to the standard brain using FLIRT [Jenkinson 2001, 2002] and then composing the functional-to-T1 and T1-to-standard registrations.

We included the co-registered COPEs corresponding to all 4 SIRP scans from each visit in a fixed-effects analysis. Finally, we performed higher-level, random-effects analyses to combine the visit-level COPEs of individual subjects.

To obtain Z-statistic maps for the activation indices described in the following, we thresholded the visit-level statistical maps using cluster-based correction for multiple comparisons [Worsley 1992]. Specifically we used a significance threshold of Z=2.3 on the cluster magnitude and P=0.05 on the cluster size.

(iii) Anatomical ROIs—We produced several anatomically defined ROIs on the left and right hemisphere for each subject. We present results here from the following ROIs:

0) Mid-temporal gyrus (MT)

1) Dorsolateral prefrontal cortex (DLPFC)

2) Dorsolateral premotor cortex (DLPMC)

3) Pre-supplementary motor cortex (PSM)

4) Supplementary motor cortex (SM)

5) Primary motor cortex in the hand region (PM)

6) Primary sensory cortex in the hand region (PS)

7) Intraparietal sulcus (IPS)

8) Insula (INS)

We included ROIs 1-8 to investigate the dependence of their activation on the memory load of the SIRP across different sites. We included ROI 0 as a control, since it is not part of the working-memory network and we did not expect its activation to be load-dependent.

To define the ROIs, we applied the FreeSurfer surface reconstruction software [Fischl 2002] on each subject's high-resolution T1 image. Conservative Talairach criteria from [Rajkowska 1995] were used to define the DLPFC. For more details on how we used the FreeSurfer subcortical segmentation and cortical parcellation to construct the ROIs for this study, we refer the reader to the supplemental material. The nine ROIs for one individual are shown in Figure 1, painted on the individual's inflated cortical surface, as obtained from FreeSurfer.

(iv) Activation indices—We obtained indices of activation for each anatomical ROI using the COPEs obtained from all four scans in a single visit (second-level fixed-effects analysis). We applied a functional mask, based on a thresholded Z-statistic, to each of the anatomical ROIs listed in the previous section. For each of these functionally masked anatomical ROIs and each of the COPEs *(i) – (iv)* listed above we extracted the following measures:

- The average percent signal change (Avg%Δ) and maximum percent signal change (Max%Δ), defined respectively as the average and maximum of each COPE normalized by the mean image intensity within each of the functionally masked ROIs. The functional mask used in this case consisted of voxels where the corrected Z-statistic for the COPE *(v)*, *i.e.*, the union of all three loads versus fixation, exceeded a threshold of Z=2.3 (cluster-corrected). Thus the exact same voxels were used to mask all other COPEs for the calculations of Avg%Δ and Max%Δ.
- The number of activated voxels (NVox), defined as the number of voxels where the corrected Z-statistic for each COPE exceeded a threshold of Z=2.3 (cluster-corrected). Thus, unlike the calculations of Avg%Δ and Max%Δ, each COPE had its own functional mask for the NVox calculations. (Using the same mask for all COPEs would have yielded the same NVox value.)

(v) Analysis of variance—The Avg%Δ, Max%Δ, and NVox measures were analyzed by mixed-model ANOVA using restricted maximum likelihood (REML), as implemented in the "lme4" package of the R statistical analysis software (http://www.r-project.org). We modeled visit (2 levels) as a fixed effect and subject (10 levels), site (4 levels), and their

interaction as random effects. We also modeled the slope of each activation measure versus memory load as a random effect, nested within subject. Formally, the activation measure $\hat{a}_{ijk}(l)$ obtained for subject $i = 1,\ldots,10$, site $j = 1,\ldots,4$, visit $k = 1,2$, and memory load $l = 1,3,5$ was modeled as

$$\widehat{a}_{ijk}(l)=a_0+s_i+\lambda_i \cdot l+t_j+u_{ij}+v_k+\varepsilon_{ijkl}, \quad (1)$$

$$s_i \sim \mathcal{N}(0,\sigma_s^2), \lambda_i \sim \mathcal{N}(\overline{\lambda},\sigma_\lambda^2), t_j \sim \mathcal{N}(0,\sigma_t^2), u_{ij} \sim \mathcal{N}(0,\sigma_u^2), \varepsilon_{ijkl} \sim \mathcal{N}(0,\sigma_\varepsilon^2)$$

where a_0 the fixed intercept, s_i the random subject effect, λ_i the random subject-dependent slope versus memory load, t_j the random site effect, u_{ij} the random subject-by-site interaction, v_k the fixed visit effect, and ε_{ijkl} the random residual. The effects were assumed independent, except for the subject-dependent terms s_i and λ_i, which were assumed to have covariance $\sigma_{s,\lambda}$. Thus the parameters estimated by the ANOVA were $\left\{a_0, v_k, \overline{\lambda}, \sigma_s^2, \sigma_\lambda^2, \sigma_{s,\lambda}, \sigma_t^2, \sigma_u^2, \sigma_\varepsilon^2\right\}$. The total variance of the activation measure at memory load l is equal to

$$\sigma_a^2(l)=\sigma_s^2+l^2\sigma_\lambda^2+2l\sigma_{s,\lambda}+\sigma_t^2+\sigma_u^2+\sigma_\varepsilon^2. \quad (2)$$

The sum $\sigma_s^2+l^2\sigma_\lambda^2+2l\sigma_{s,\lambda}$ can be seen as the overall variance due to subject variability, including variability in the magnitude and slope of the subjects' response.

(vi) Power analysis—The differences in how activation measures vary as a function of memory load in patients versus controls are used in schizophrenia research to characterize working memory deficits in patients [Manoach 1999, Manoach 2000, Ragland 2007]. Here we calculate the power for a test on the difference of the mean activation-versus-load slopes of two populations.

Specifically, we calculate the power for a hypothetical multi-site study, where each subject is scanned once at a single site. We then use the variance estimates derived from the present study to establish how power is affected when we pool subjects from different sites. This allows us to determine if pooling data from several sites leads to greater power because of a greater total number of subjects or if pooling data from sites with different levels of noise negates the benefit of increasing the number of subjects. This analysis mirrors the one from [Suckling 2008], except that the effect of interest and the model of variance components differ in our case.

Let $a(l)$, $a'(l)$ be the true activation measure for the two populations of interest as a function of memory load and λ, λ' the respective slopes. The effect that we want to test is the difference of slopes, given by

$$\Delta\lambda=\lambda-\lambda'=\frac{a(l_2)-a(l_1)}{l_2-l_1}-\frac{a'(l_2)-a'(l_1)}{l_2-l_1} \quad (3)$$

for any two memory loads l_1, l_2. We estimate the effect from noisy observations of the activation measure, $\widehat{a}_i(l),\ i=1,\ldots,N,\ \widehat{a}'_i(l),\ i=1,\ldots,N'$, where N, N' the numbers of subjects drawn from the two populations. We assume that each of the subjects in the study is

scanned at one site only and for one visit only, thus we drop the dependence of the activation on the site and visit here. Different subjects, however, may be scanned at different sites. Then the estimated effect is

$$\Delta\widehat{\lambda} = \frac{1}{l_2 - l_1}\left\{\frac{1}{N}\sum_{i=1}^{N}[\widehat{a}_i(l_2) - \widehat{a}_i(l_1)] - \frac{1}{N'}\sum_{i=1}^{N'}[\widehat{a}'_i(l_2) - \widehat{a}'_i(l_1)]\right\}. \quad (4)$$

Substituting from (1) yields

$$\Delta\widehat{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_i + \frac{\varepsilon_{il_2} - \varepsilon_{il_1}}{l_2 - l_1}\right) + \frac{1}{N'}\sum_{i=1}^{N'}\left(\lambda'_i + \frac{\varepsilon'_{il_2} - \varepsilon'_{il_1}}{l_2 - l_1}\right), \quad (5)$$

where λ_i, λ'_i the slopes for the i-th subject in each of the two populations and ε_{il}, ε'_{il} the noise for the i-th subject at memory load l.

Let σ_λ^2, $\sigma_{\lambda'}^2$ be the variance of the random slope in the two populations and σ_ε^2 the variance of the random noise in the activation measurements. Then, assuming independence, the variance of the estimated effect from (5) is

$$\mathrm{Var}\{\Delta\widehat{\lambda}\} = \frac{\sigma_\lambda^2}{N} + \frac{\sigma_{\lambda'}^2}{N'} + \frac{2\sigma_\varepsilon^2}{(l_2 - l_1)^2}\left(\frac{1}{N} + \frac{1}{N'}\right). \quad (6)$$

The power of the two-sided test between hypotheses H_1 :$|\Delta\lambda| > 0$ and H_0:$|\Delta\lambda| = 0$ is

$$\mathrm{Power} = \frac{1}{2}\mathrm{erfc}\left(\frac{T - \mathrm{E}[\Delta\widehat{\lambda}]}{\sqrt{2\mathrm{Var}\left\{\Delta\widehat{\lambda}\right\}}}\right) + \frac{1}{2}\mathrm{erfc}\left(\frac{T + \mathrm{E}[\Delta\widehat{\lambda}]}{\sqrt{2\mathrm{Var}\left\{\Delta\widehat{\lambda}\right\}}}\right), \quad (7)$$

where erfc(·) is the complimentary error function and T the detection threshold.

For the power analysis presented in the following we used expected effect sizes $\mathrm{E}[\Delta\widehat{\lambda}]$ for the left and right DLPFC from a previous study involving 9 schizophrenia patients and 9 healthy controls [Manoach 2000]. That study used the same WM paradigm but somewhat different fMRI data acquisition parameters (1.5T field strength, in-plane resolution 3.13mm, slice thickness 8 mm, TE=50msec, TR=2sec, FA=70°) than the present study. Thus the validity of our results depends on the generalizability of the findings of [Manoach 2000].

We calculated the effect variance $\mathrm{Var}\left\{\Delta\widehat{\lambda}\right\}$ from (6), using estimates of the slope and noise variances from the variance components analysis of our reliability data set. This allowed us to calculate power for a scenario similar to the clinical schizophrenia study of the MCIC, which involved the same sites as the present study but large numbers of patients and healthy controls ($N \approx N' \approx 150$), each scanned at one of the four sites. For simplicity and to sum up to a total number of patients and controls similar to that study, we assumed in our calculations that each site contributed equally, with 38 patients and 38 controls. We set the detection threshold T to achieve type-I error probability equal to 0.05.

3. Results

I. Behavioral data

All subjects performed at or near ceiling levels of overall accuracy (range: 86-99% correct, mean: 95%±4% correct). There was little decrease in average accuracy as the memory load increased: 96%±3% (load 1), 95%±3% (load 3), 92%±5% (load 5). Each subject visited the four sites in a different order. Figure 2 shows a plot of each subject's accuracy at the four sites in the order they were visited. Accuracy was stable over time for most subjects, except for two that exhibited a decrease in accuracy with time. Thus there was no evidence of learning effects.

Figure 3 shows plots of the average RT versus memory load for each of the four sites with standard error bars, illustrating that RT and its slope as a function of memory load was reliable across sites. Although the RTs recorded at Iowa appear to be somewhat longer than those recorded at other sites, the differences between the slope and intercept of the RT vs. memory load as estimated at different sites were not statistically significant. Specifically, two-sided unpaired T-tests between sites yielded p-values of 0.37 or higher for the slope differences and 0.27 or higher for the intercept differences.

The analysis of variance that we performed on the RT data indicated that the variability between subjects was much higher than the variability between sites. Specifically, the variances attributed by the ANOVA to each of the random effects, expressed as percentages of the overall RT variance, were as follows: subject 15.8%, site 1%, site-by-subject interaction 2.9%, run 1.5%, and residual 78.9%. The fixed effects were estimated as follows: intercept 515.9ms, memory load 32.8ms, probe type (foil–target) 38.6ms, visit (2^{nd}–1^{st}) -9.8ms, and site order (2^{nd}–1^{st}/3^{rd}–1^{st}/4^{th}– 1^{st}) 2.7/-2.2/-8.6ms. These results demonstrate reasonable stability of performance over time and across sites. Consequently, we do not expect learning effects or other behavioral variability to confound our fMRI data analysis.

II. fMRI data

As an example of how fMRI activation varied across subjects and sites, Figure 4 shows plots of Avg%Δ for three ROIs in the left hemisphere. Two of the ROIs displayed here (DLPFC, IPS) are typically hypothesized to be part of the working-memory network, while the third ROI (MT) is not. Thus the first two ROIs are expected to exhibit activation dependent on the memory load but the third is not. The bars in Figure 4 represent Avg%Δ for each of the three memory loads, shown with standard error bars. We show averages of Avg%Δ over all data, over each site, and over each subject. We also mark the cases that exhibit significance or trend towards significance on the difference between Avg%Δ for load 5 and load 1 based on a two-sided T-test.

These plots illustrate the general trends in the data. Specifically, subjects were found to differ both in their magnitude of activation indices, and in the dependence of activation indices on memory load. Although there was across-site variability of the indices, it was smaller than the across-subject variability. General trends were captured in the results of all four sites. For example, dependence on memory load was more significant in the left IPS than the left DLPFC at all sites. Although a small increase in activation occurred at the highest load compared to the lowest load in the control region of MT, the load dependence in that area was not significant in any of the sites.

Figure 5 shows plots of Max%Δ by memory load for the same three ROIs as above, averaged over all data sets acquired at each field strength. The plots show a modest increase in activation at 3T in the ROIs that are hypothesized to be part of the working-memory network (DLPFC, IPS) and a modest decrease in activation at 3T in the control ROI (MT).

The contribution of site and subject to the variability in the data is further quantified in Figure 6 for all 9 ROIs and all 3 activation measures studied here. The plots show the variance attributed by the mixed-model ANOVA to the factors of subject $\left(\sigma_s^2+l^2\sigma_\lambda^2+2l\sigma_{s,\lambda}\right)$, site $\left(\sigma_t^2\right)$ and subject-by-site interaction $\left(\sigma_u^2\right)$, as well as the residual variance $\left(\sigma_\varepsilon^2\right)$, as percentages of the total variance, for Avg%Δ, Max%Δ, and NVox. Note that, because the subject variance is a function of the memory load *l*, its contribution to the overall variance differs at each load. The plots in Figure 6 were generated for $l = 3$. The variance components for the COPEs of each of the 3 memory loads versus fixation are summarized in Table 1, which shows median percentages of variance over all 18 ROIs (9 in each hemisphere).

At the end of Table 1 we also show results from a separate mixed-model ANOVA that we applied to the COPE of load 5 versus load 1. The model for this particular ANOVA was similar to the one used for the COPEs of individual loads vs. fixation from equation (1), except without the load dependent term $\lambda_i \cdot l$. Other than the result from the COPE of load 5 versus load 1 at the bottom of Table 1, all other variance components results reported in the following refer to the model of equation (1), which was applied to the COPEs of individual loads versus fixation.

The three activation measures that we studied exhibited differences in terms of their variance components. The percentage of Avg%Δ variance that was attributed to the site factor was an order of magnitude smaller than the percentages attributed to the subject and subject-by-site factors. The percentage of variance attributed to site was greater for Max%Δ and NVox than it was for Avg%Δ. However, even for those measures, it was smaller than the percentages contributed by subject and subject-by-site interaction. The contribution of the subject-by-site interaction suggests that variability between data acquired at different sites was due more to individual subjects activating differently on different occasions than to overall site differences. Variance components corresponding to each of the three memory load versus fixation were fairly similar to each other for Avg%Δ and Max%Δ, but not so for NVox. Because the variance of load slopes over subjects, σ_λ^2, was more substantial for NVox than the other two activation measures, the dependence of the subject-related variance, $\sigma_s^2+l^2\sigma_\lambda^2+2l\sigma_{s,\lambda}$, on the memory load was greater for NVox as well.

Table 2 shows the relative contributions of the factors comprising subject variability, summarized through their median values over all 18 ROIs. Both the slope variance and the covariance of slope and intercept were found to be one to two orders of magnitude smaller than the intercept variance. These results confirm that the contribution of the slope variance σ_λ^2 was greater for NVox than it was for Avg%Δ and Max%Δ.

A strategy suggested in [Friedman 2008] for evaluating across-site reliability by identifying outlier sites is to repeat analyses after removing each one of the sites consecutively. Following this strategy, we repeated the ANOVA four times, excluding the data of one of the four sites and including the other 3 sites each time. In the following we compare ANOVA results obtained after each site was excluded. Since it would be unclear how to choose absolute thresholds on these ANOVA outcomes to deem sites as outliers, we did not compare the results obtained by excluding each site to any absolute threshold, but only relative to the results from the exclusion of the other sites and the results obtained when all four sites were included.

Table 3 shows how the percentage of the overall variance that was attributed to the site factor changed as we excluded each of the sites. The table shows median changes over all 18

ROIs. A positive change indicates that the percentage of variance attributed to site increased when the specific site was excluded. If there were a dramatic decrease of site-by-site variance over all activation measures when a specific site was removed, it would suggest that this site was an outlier. None of the sites emerged as an outlier based on this criterion. Specifically, there was no site whose exclusion led to a decrease in the variance of all 3 activation measures. Among activation measures, Avg%Δ was the most stable in this comparison, with only small increases of variance (less than 10% of the corresponding across-site variance from Table 1) when any single site was removed.

Figure 7 shows plots of the grand mean of the slope of activation versus memory load over all subjects, $\bar{\lambda}$, as estimated by the ANOVA, with standard error bars. We also applied the strategy of removing one site to determine how the slope estimate changed when each site's data was excluded. Table 4 shows that the slopes of activation versus memory load that we obtain from any combination of three of the sites are very similar to each other and very similar to those obtained by combining all four sites. We quantified differences in the slope estimates via the normalized mean squared error, $\sqrt{\sum_i \left[\bar{\lambda}_3(i) - \bar{\lambda}_4(i)\right]^2} / \sqrt{\sum_i \bar{\lambda}_4(i)^2}$, where $\bar{\lambda}_3(i)$ and $\bar{\lambda}_4(i)$ are the mean slopes for the i-th ROI when including 3 sites and all 4 sites respectively. None of the sites emerged as a clear outlier from this comparison either, as would have been indicated by a much greater change in the slope estimates for all 3 measures when removing the specific site. For all ROIs and all activation measures, the differences between the slopes estimated from 3 sites and the slope estimated from all 4 sites were within the standard errors shown in Figure 7.

Table 5 shows the results of the power analysis. We had at our disposal estimates of activation measures in the left and right DLPFC in schizophrenia patients and healthy controls from a previous study [Manoach 2000]. We used those estimates (for Avg%Δ in load 5 versus fixation and load 2 versus fixation) to obtain our expected effect size, *i.e.*, the expected difference of activation slopes between patients and controls (in units of percentage points by memory target digit). The effect size was 0.003 for the left DLPFC and 0.013 for the right DLPFC. We substituted this effect size into equation (7) to calculate the power. We repeated the calculation with the variances obtained from the ANOVA when the data from all four sites was included, when data from one of the sites was excluded, and when only data from a single site was included. As seen in Table 5, power would be greatest if all sites were included. Thus, the advantage of pooling more subjects would outweigh any potential disadvantage from differences in the pooled sites' levels of noise.

Equations (6) and (7) can also be used to calculate the sample size required to achieve a certain power. In general, scanning all subjects at the site with the lowest data variance would require the smallest sample size. Distributing the subjects over multiple sites would require a sample size greater than what would be needed if all subjects were scanned at the site with lowest variance but smaller than what would be needed if all subjects were scanned at the site with highest variance. Thus in the multi-site setting we are interested in the trade-off between scanning all subjects at the single site where power was highest, thus reducing the total number of subjects required, versus scanning subjects at multiple sites and parallelizing the scans, thus reducing the overall time required to complete the study.

To quantify this trade-off we fixed the power levels for left and right DLPFC at the levels achieved by using data from all four sites and $N = N' = 150$ patients and controls. We then used the variance estimated from the ANOVA on each individual site to calculate how many subjects would be needed to achieve the same power if all subjects were scanned at that site. The results are shown in Table 6. Because the ranking of the sites with respect to their individual power at a fixed sample size was different in the left and right DLPFC (see Table

5), the site ranking with respect to the required sampled size at a fixed power level was also different in the two ROIs. If one wanted to achieve the desired power levels in both ROIs, one would pick the highest of the sample sizes required for each ROI. Thus, the required sample sizes $N = N'$ at each site would be 162 (UNM), 171 (Iowa), 164 (MGH), and 135 (Minnesota). Minnesota was the only site where there would be savings in the required sample size compared to the 150 we would need to reach the same power if we distributed the scans across the four sites. Those savings would be gained at the expense of longer times needed to recruit and scan subjects at a single site, particularly for large studies such as the one in this example. For smaller studies, on the other hand, the additional time required to set up multi-site data collection properly might trump the time saved by parallelizing the data collection.

Finally, for a visual comparison of the similarity of statistical maps among subjects and sites, we show Z-maps for the contrast of highest versus lowest memory load (load 5 versus load 1) on the left hemisphere. The maps have been thresholded to show the top 2% of Z-values and mapped to the inflated surface reconstructed from the T1 images of one of the subjects. The anatomical ROIs are shown overlaid on that surface as well. Specifically, Figure 8 shows individual maps for each of the ten subjects, each obtained from a fixed-effects analysis of one subject's scans at all four sites. Figure 9 shows maps for each of the four sites, each obtained from a random-effects analysis of scans from all ten subjects acquired at one site. In Figure 8 each subject's statistical maps are shown on the subject's own anatomy. In Figure 9 all maps are shown on the anatomy of one of the subjects.

These statistical maps support the established observation that individual subjects differ from one another not only in the magnitude and load dependence of their activations but also in the spatial localization of their activation peaks. Note that the top 2% of load-dependent activations, as shown in these figures, correspond to different thresholds in each case. This is not the usual way that one would choose the threshold, *e.g.*, to compare statistical significance of activation magnitudes between different maps. However, these figures are intended as a comparison of the *locations* of activation peaks, not as a comparison of the activation magnitudes.

The activation peaks obtained by the group analysis of the ten subjects' data from each site, shown in Figure 9, appear less variable in their location than the ones obtained for individual subjects, shown in Figure 8. For quantitative measures of similarity of the whole-brain statistical maps between sites and between subjects we refer the reader to the supplemental material.

4. Discussion

Our study indicates that it is possible to obtain fMRI activation indices for working memory processing that are reliable across sites with different scanners. Thus it is possible to combine multi-site data to improve power in studies of such a task. However, our results also show that it is important to choose an appropriate activation measure. In particular, the average percent signal change was the most reliable among the measures we studied, with its across-site variability being an order of magnitude smaller than its across-subject variability. The maximum percent signal change was somewhat less reliable across sites, and the number of activated voxels was the least reliable. In addition, the number of activated voxels exhibited greater variability among subjects in terms of its slope versus memory load, leading to a greater dependence of its variance components on the memory load.

It has been known that different activation measures can yield different degrees of variability. Some of the activation measures that have been used to assess across-site reliability are percent signal change and volume of activation [Costafreda 2007, Sutton

2008], which are threshold-dependent, or F-statistic values [Suckling 2008], which are threshold-independent. Friedman *et al.* showed that measures based on image contrast (median and maximum percent signal change) are more reliable than measures based on contrast-to-noise ratio, possibly because the latter involve an estimate of variance in the denominator, which can be difficult to obtain reliably [Friedman 2008]. Thus we have focused here mainly on contrast-based measures, namely the average and maximum percent signal change. We expect the median measure to lie between the average and maximum measures in terms of reliability. We found the number of activated voxels to be less reliable than contrast-based measures, perhaps because of its dependence on both the mean and the variance of the signal. This was by no means an exhaustive comparison and further evaluation of the reliability of different activation measures in working-memory processing is warranted.

Another aspect of data processing that could affect reliability is the type of ROIs used to extract activation measures. When prior hypotheses cannot be formulated on a purely anatomical basis, it is common to apply functional masks to anatomically defined ROIs. A recent comparison of strategies for functional mask definition has shown that using masks derived from individual conditions introduces unacceptable biases in average signal change calculations, which are not present when masks are derived from the union of all relevant conditions [Mitsis 2008]. In keeping with this we have chosen to use functional masks derived from the union of all three memory loads for our average and maximum signal change calculations.

In this work we chose to use functional masks derived from the data of each visit separately (second-level fixed-effects analysis). An alternative approach is to derive a common functional mask from the data of both visits (third-level fixed-effects analysis). We repeated our ANOVA with third-level masks and the results are summarized in the supplemental material. As with second-level masks, the variance due to site was much smaller than the other variance components.

The threshold applied to the functional masks is another choice to consider. As the threshold becomes more stringent, the size of the functional mask decreases and thus the average and the maximum signal converge, eliminating the advantage of the former over the latter in terms of reliability. The minimum size of clusters surviving cluster-based thresholding depends on the autocorrelation function of the images [Forman 1995]. For our data and the threshold we used, this minimum cluster size was 904 mm^3 (UNM), 1440 mm^3 (Iowa), 1104 mm^3 (MGH), and 1128 mm^3 (Minn).

Data processing strategies recommended in [Friedman 2008] to improve across-site reliability were: *(i) Increasing the size of the ROIs, especially for 3T data*. The average volume of the anatomical ROIs used here was 10212mm^3 ± 5717 mm^3. We believe that these ROIs, based on cortical parcellations of T1 images, were sufficiently large to ensure reliable activation measures. *(ii) Compensating for differences in the smoothness of the images acquired at different sites, especially for contrast-to-noise measures*. Images from the four sites did exhibit somewhat different smoothness. Using spatial smoothness estimation available in FreeSurfer, we calculated the average full width at half maximum for each site's images at 2.79±0.15mm (UNM), 4.03±0.36mm (Iowa), 3.13±0.10mm (MGH), and 3.14±0.57mm (Minnesota). That is, smoothness was similar for the two Siemens 3T sites, slightly lower for the Siemens 1.5T site, and higher for the GE 1.5T site. It is possible that equalizing the smoothness of the images from the four sites could increase reliability further. However, the Iowa site, which would be most affected from such an equalization, did not emerge as an outlier in our analyses. That is, as seen in Table 3 through Table 5, exclusion of the Iowa data did not lead to consistently better or worse performance compared to the

exclusion of the other sites, thus giving us no indication of a systematic bias due to smoothness differences between Iowa and the other sites. Based on this result we chose not to pursue smoothness equalization further. *(iii) Including more scans*. The maximum number of sensorimotor scans included in [Friedman 2008] was four, which equals the number of the SIRP scans included in the present study. It is possible that acquiring additional scans could improve reliability further. However, it would be impractical due to subject fatigue, particularly in studies of populations with psychiatric disorders such as schizophrenia.

Another strategy for improving across-site reliability that has been proposed in the past is to compensate for differences in signal-to-fluctuation-noise-ratio (SFNR) between the sites [Friedman 2006b]. We found the SFNR for each site in our study to be 110 (UNM), 156 (Iowa), 149 (MGH), and 149 (Minnesota). Repeating our ANOVA with SFNR as a regressor did not yield an improvement in reliability, thus we chose not to include it in the results reported here.

We did not perform any explicit correction for image distortions due to field inhomogeneities. Since inhomogeneity effects are expected to be more salient in 3T sites (MGH and Minnesota), correction based on field maps could lead to better alignment of the images from those sites, both to the individual T1 and to the standard brain images. There are two kinds of field inhomogeneity artifacts in EPI: geometric distortions of the images in the phase-encode direction and signal loss in tissue interface areas. The functional-to-anatomical registration strategy employed here could mitigate geometric distortion effects but of course it could not compensate for signal loss effects.

As discussed in the Methods section, identifying outlier scans based on image intensity and motion criteria and removing them from the analysis did not have an impact on the reliability of the estimated activation measures in this study. However, the participating subjects were highly motivated and healthy volunteers. We still believe that data quality assurance checks like these are important and may have a greater impact on studies of clinical populations, such as schizophrenia patients, where we expect to see more subject motion.

When pooling the data from all four sites, we found a positive slope of activation versus memory load in a number of regions. For example the estimated slope for Avg%Δ, as seen in Figure 7, was greatest in the left IPS, followed by left PSM, left PM, left DLPMC, and right DLPFC. Maps of load-dependent activations obtained when we pool data from all four sites are shown in the supplemental material. Note that in the left hemisphere a high activation peak occurs just posterior to the DLPFC/DLPMC border. This border was based on the Talairach coordinate suggested in [Rajkowska 1995]. A slight change in that coordinate could include that activation peak in the left DLPFC ROI and thus lead to higher load dependence in that ROI than what we report here. In addition, some areas of high load-dependent activation, such as the left inferior frontal cortex, were not among the ROIs that we studied. However, we believe that the ROIs included here represent a reasonably broad range of areas and we would not expect dramatically different results in other ROIs.

In conclusion, our study illustrates the feasibility of deriving fMRI-based measures of working-memory processing that are much more variable across subjects than they are across sites in the brain regions studied here. This is attainable even across sites with different scanner manufacturers and field strengths, as long as appropriate scanner calibration and data processing methods are used. Our results are encouraging for multi-site studies of similar paradigms, where subjects are recruited at different geographic locations to improve the generalizability of the results, to accelerate subject accrual, or to investigate conditions with low prevalence in the general population.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Costafreda SG, Brammer MJ, Vêncio RZ, Mourão ML, Portela LA, de Castro CC, Giampietro VP, Amaro E Jr. Multisite fMRI reproducibility of a motor task using identical MR systems. J Magn Reson Imaging. 2007; 26(4):1122–1126. [PubMed: 17896376]

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. Magn Res Med. 1995; 33(5):636–647.

Friedman L, Glover GH. Report on a Multicenter fMRI Quality Assurance Protocol. J Magn Reson Imaging. 2006; 23(6):827–839. [PubMed: 16649196]

Friedman L, Glover GH, The FBIRN Consortium. Reducing interscanner variability of activation in a multicenter fMRI study: Controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage. 2006; 33(2):471–481. [PubMed: 16952468]

Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, Greve DN, Bockholt HJ, Belger A, Mueller B, Doty MJ, He J, Wells W, Smyth P, Pieper S, Kim S, Kubicki M, Vangel M, Potkin SG. Test-retest and between-site reliability in a multicenter fMRI study. Human Brain Mapping. 2008; 29(8):958–972. [PubMed: 17636563]

Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. Magn Res Med. 1996; 35(3):346–355.

Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. NeuroImage. 2009; 48(1):63–72. [PubMed: 19573611]

Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pachec J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale A, Dickerson B, Fischl B. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage. 2006; 32(1):180–194. [PubMed: 16651008]

Jenkinson M, Smith SM. A Global Optimisation Method for Robust Affine Registration of Brain Images. Medical Image Analysis. 2001; 5(2):143–156. [PubMed: 11516708]

Jenkinson M, Bannister P, Brady M, Smith S. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. NeuroImage. 2002; 17(2):825–841. [PubMed: 12377157]

Jovicich J, Czanner S, Han X, Salat D, van der Kouwe A, Quinn B, Pacheco J, Albert M, Killiany R, Blacker D, Maguire P, Rosas D, Makris N, Gollub R, Dale A, Dickerson BC, Fischl B. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. Neuroimage. 2009; 46(1):177–192. [PubMed: 19233293]

Kristofferson MW. Effects of practice on character-classification performance. Canad J Psychol/Rev Canad Psychol. 1972; 26:54–60.

Manoach DS, Schlaug G, Siewert B, Darby DG, Bly BM, Benfield A, Edelman RR, Warach S. Prefrontal cortex fMRI signal changes are correlated with working memory load. NeuroReport. 1997; 8(2):545–549. [PubMed: 9080445]

Manoach DS, Press DZ, Thangaraj V, Searl MM, Goff DC, Halpern E, Saper CB, Warach S. Schizophrenic subjects activate dorsolateral prefrontal cortex during a working memory task as measured by fMRI. Biol Psychiatry. 1999; 45(9):1128–1137. [PubMed: 10331104]

Manoach DS, Gollub RL, Benson ES, Searl MM, Goff DC, Halpern E, Saper CB, Rauch SL. Schizophrenic subjects show aberrant fMRI activation of dorsolateral prefrontal cortex and basal

ganglia during working memory performance. Biol Psychiatry. 2000; 48(2):99–109. [PubMed: 10903406]

Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, Kennedy DN, Gollub RL. Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am J of Psychiatry. 2001; 158(6):955–958. [PubMed: 11384907]

Mitsis GD, Iannetti GD, Smart TS, Tracey I, Wise RG. Regions of interest analysis in pharmacological fMRI: How do the definition criteria influence the inferred result? NeuroImage. 2008; 40(1):121–132. [PubMed: 18226552]

Ragland JD, Yoon J, Minzenberg MJ, Carter CS. Neuroimaging of cognitive disability in schizophrenia: Search for a pathophysiological mechanism. Int Rev of Psychiatry. 2007; 19(4): 419–429.

Rajkowska G, Goldman-Rakic PS. Cytoarchitectonic Definition of Prefrontal Areas in the Normal Human Cortex: I. Remapping of Areas 9 and 46 using Quantitative Criteria. Cerebral Cortex. 1995; 5:307–322. [PubMed: 7580124]

Smith S. Fast Robust Automated Brain Extraction. Human Brain Mapping. 2002; 17(3):143–155. [PubMed: 12391568]

Sternberg S. High-speed scanning in human memory. Science. 1966; 153:652–654. [PubMed: 5939936]

Suckling J, Ohissen D, Andrew C, Johnson G, Williams S, Graves M, Chen CH, Spiegelhalter D, Bullmore E. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Human Brain Mapping. 2008; 29(10):1111–1122. [PubMed: 17680602]

Sutton BP, Goh J, Hebrank A, Welsh RC, Chee MW, Park DC. Investigation and validation of intersite fMRI studies using the same imaging hardware. J Magn Reson Imaging. 2008; 28(1):21–28. [PubMed: 18581342]

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

Whitfield-Gabrieli, S. Artifact Detection Tools. 2009. http://web.mit.edu/swg/software.htm

Woolrich MW, Ripley BD, Brady JM, Smith SM. Temporal Autocorrelation in Univariate Linear Modelling of FMRI Data. NeuroImage. 2001; 14(6):1370–1386. [PubMed: 11707093]

Worsley KJ, Evans AC, Marrett S, Neelin P. A three-dimensional statistical analysis for CBF activation studies in human brain. Journal of Cerebral Blood Flow and Metabolism. 1992; 12:900–918. [PubMed: 1400644]

Zou KH, Greve DN, Wang M, Pieper SD, Warfield SK, White NS, Manandhar S, Brown GG, Vangel MG, Kikinis R, Wells WM III, FIRST BIRN. Reproducibility of Functional MR Imaging: Preliminary Results of a Prospective Multi-institutional Study Performed by the Biomedical Informatics Research Network. Radiology. 2005; 237:781–789. [PubMed: 16304101]

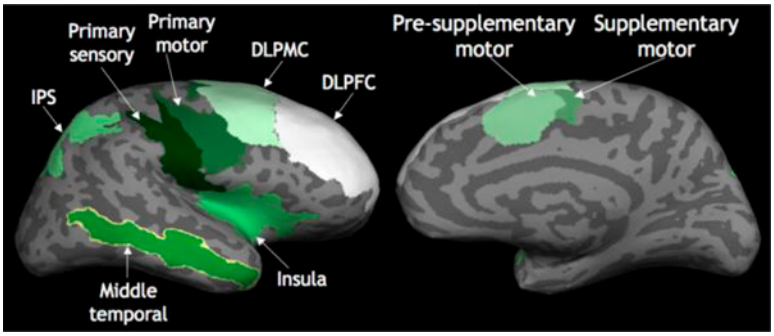

Figure 1. Anatomical ROIs shown on the inflated cortical surface of an individual

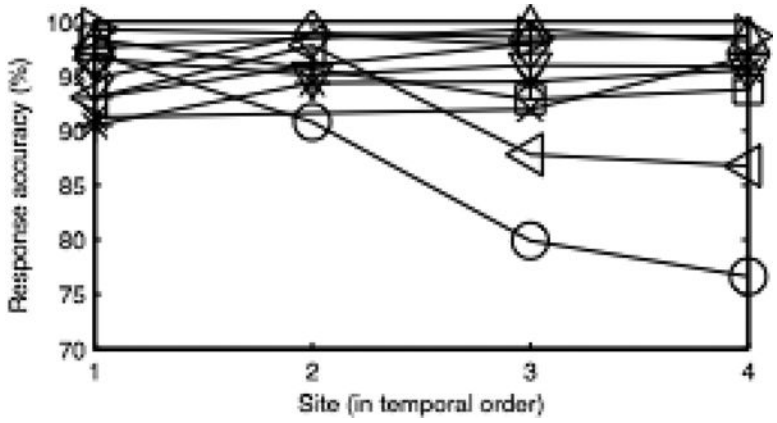

Figure 2. Response accuracy at each site in the order the sites were visited. Each plot corresponds to a different subject

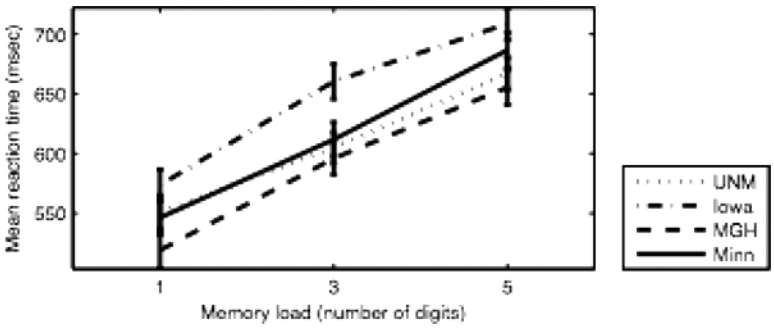

Figure 3. Reaction time versus memory load, averaged over all scans for each site

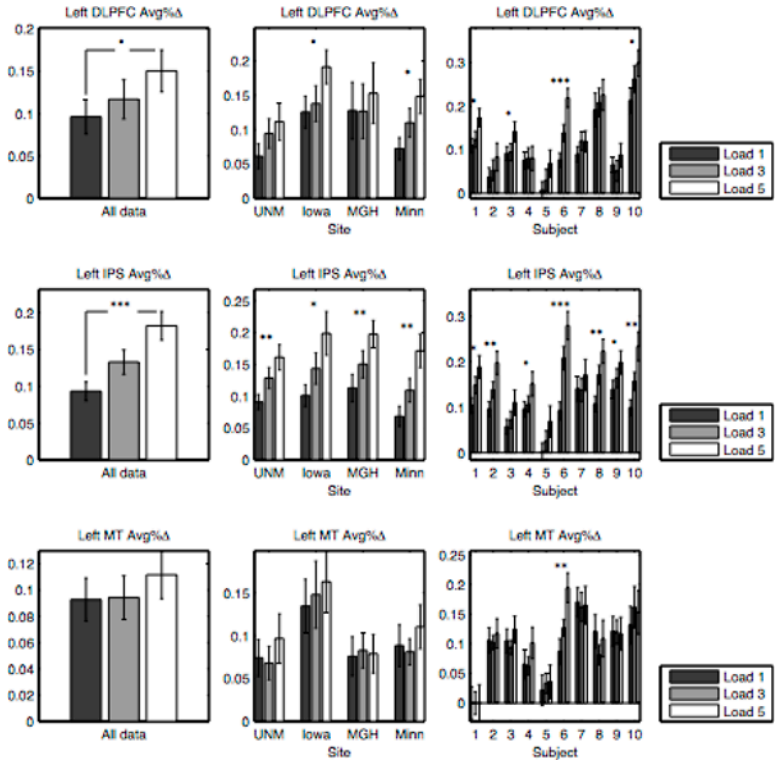

Figure 4. Activation (Avg%Δ) by memory load in three ROIs of the left hemisphere. From left to right: average activation over all data, by site, and by subject. P-values on the difference between load 5 and load 1 have been marked as follows; p<.1: •; p<.05: *; p<.01: **; p<.001: *

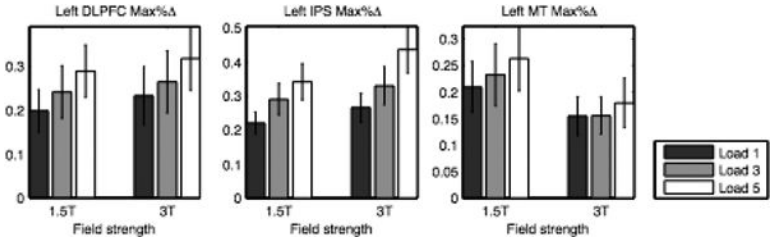

Figure 5. Activation (Max%Δ) by memory load and field strength in three ROIs of the left hemisphere

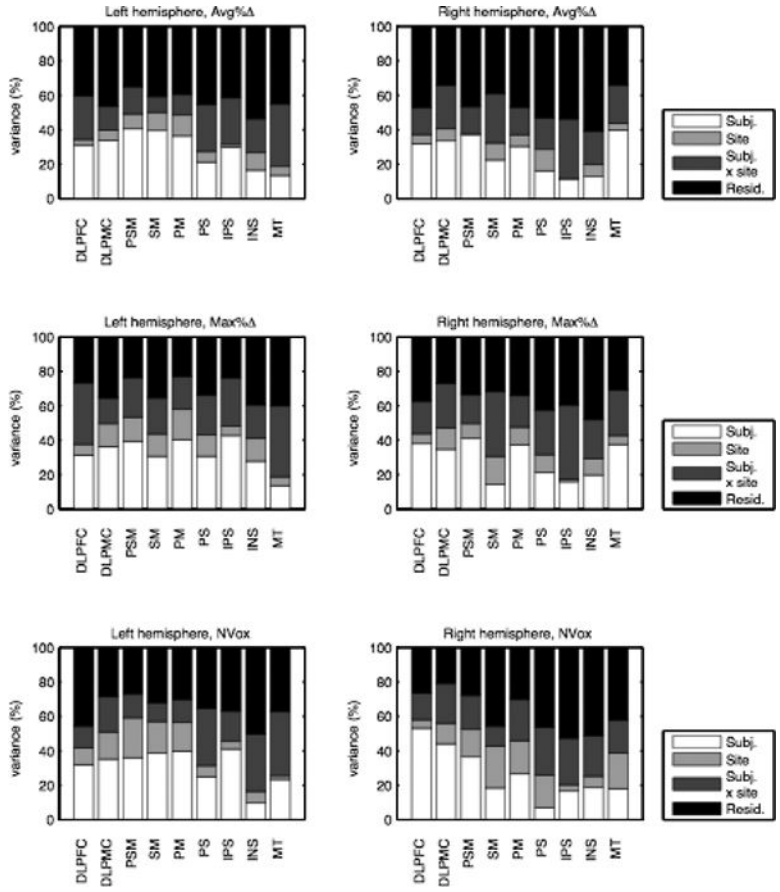

Figure 6. Variance components (subject $\sigma_s^2+l^2\sigma_\lambda^2+2l\sigma_{s,\lambda}, l=3$, site σ_t^2, subject-by-site interaction σ_u^2, Residual σ_ε^2) of Avg%Δ, Max%Δ, and NVox for every ROI in the left and right hemispheres

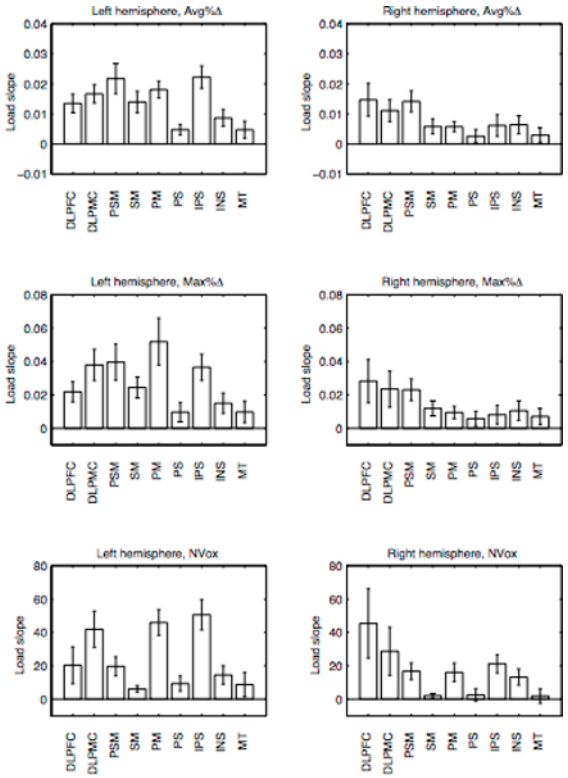

Figure 7. Estimated slope of Avg%Δ, Max%Δ, and NVox versus memory load in the left and right hemispheres, using all data

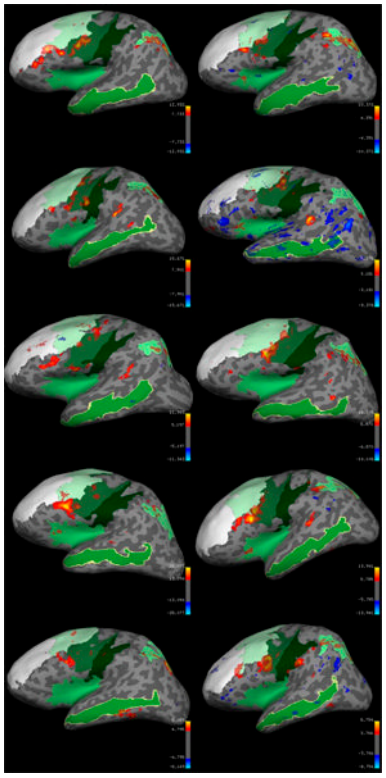

Figure 8. Top 2% of z-values for contrast of highest versus lowest memory load, shown for each of the 10 subjects

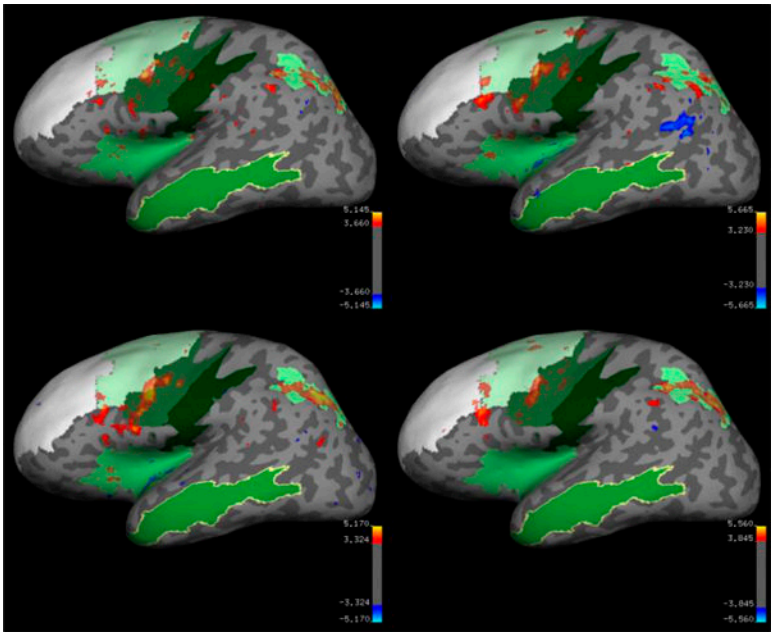

Figure 9. Top 2% of z-values for contrast of highest versus lowest memory load, shown for each of the 4 sites. (Clockwise from top left: UNM, Iowa, Minnesota, MGH.)

Table 1

Median percentage of variance attributed to subject, site, and subject-by-site interaction for activation measures derived from each COPE

		Avg%Δ	Max%Δ	NVox
Load 1 vs. fixation	**Subject**	25%	26%	19%
	Site	7%	11%	18%
	Subject x Site	21%	25%	25%
Load 3 vs. fixation	**Subject**	31%	33%	29%
	Site	7%	10%	14%
	Subject x Site	18%	23%	20%
Load 5 vs. fixation	**Subject**	36%	39%	43%
	Site	6%	10%	10%
	Subject x Site	18%	21%	16%
Load 5 vs. load 1	**Subject**	18%	33%	36%
	Site	3%	5%	4%
	Subject x Site	4%	9%	19%

Table 2

Median relative magnitude of subject-dependent components of variance: intercept variance σ_s^2, slope variance σ_λ^2, and covariance of intercept and slope $\sigma_{s,\lambda}$

	Avg%Δ	Max%Δ	NVox
$\sigma_\lambda^2/\sigma_s^2$	0.0335	0.0357	0.1792
$\lvert\sigma_{s,\lambda}\rvert/\sigma_s^2$	0.0760	0.0647	0.0970

Table 3

Median change in the percentage of overall variance due to site, when one of the sites is excluded from the analysis (calculated at memory load 3)

Excluded site	Avg%Δ	Max%Δ	NVox
UNM	+0.1%	-2.7%	-3.7%
Iowa	+0.2%	+0.2%	-5.0%
MGH	+0.3%	+1.9%	+5.7%
Minn	+0.6%	+3.6%	+2.9%

Table 4

Normalized mean squared error between slopes of activation versus memory load estimated with one site excluded and those estimated from all four sites

Excluded site	Avg%Δ	Max%Δ	NVox
UNM	8%	11%	7%
Iowa	9%	14%	9%
MGH	15%	12%	18%
Minn	14%	12%	15%

Table 5

Power of a T-test on group differences in the slope of Avg%Δ versus memory load

ROI	All 4 sites included	Exclude one site				Single site only			
		UNM	Iowa	MGH	Minn	UNM	Iowa	MGH	Minn
Left DLPFC	0.14	0.12	0.11	0.12	0.12	0.07	0.08	0.07	0.07
Right DLPFC	0.69	0.54	0.61	0.59	0.56	0.31	0.21	0.22	0.28

Neuroimage. Author manuscript; available in PMC 2013 October 28.

Table 6

Sample size ($N = N'$) required to achieve equal power for a T-test on group differences in the slope of Avg%Δ versus memory load, using data from all four sites versus using data from a single site only

ROI	All 4 sites	Single site only			
		UNM	Iowa	MGH	Minn
Left DLPFC	150	162	113	146	135
Right DLPFC	150	105	171	164	122