Original Research

Investigation and Validation of Intersite fMRI Studies Using the Same Imaging Hardware

Bradley P. Sutton, PhD,[1,2]* Joshua Goh, MA,[2–4] Andrew Hebrank, BA,[2,5] Robert C. Welsh, PhD,[6] Michael W.L. Chee, MD,[4] and Denise C. Park, PhD[2,3,5]

Purpose: To provide a between-site comparison of functional MRI (fMRI) signal reproducibility in two laboratories equipped with identical imaging hardware and software. Many studies have looked at within-subject reliability and more recent efforts have begun to calibrate responses across sites, magnetic field strengths, and software. By comparing identical imaging hardware and software, we provide a benchmark for future multisite comparisons.

Materials and Methods: We evaluated system compatibility based on noise and stability properties of phantom scans and contrast estimates from repeated runs of a blocked motor and visual task on the same four subjects at both sites.

Results: Analysis of variance (ANOVA) and region of interest (ROI) analysis confirmed that site did not play a significant role in explaining variance in our large fMRI dataset. Effect size analysis shows that between-subject differences account for nearly 10 times more variance than site effects.

Conclusion: We show that quantitative comparisons of contrast estimates derived from cognitive experiments can reliably be compared across two sites. This allows us to establish an effective platform for comparing group differences between two sites using fMRI when group effects are potentially confounded with site, as in the study of neurocultural differences between countries or multicenter clinical trials.

Key Words: functional MRI; reproducibility; intersite comparisons; effect size; cultural neuroscience
J. Magn. Reson. Imaging 2008;28:21–28.

SINCE THE INTRODUCTION of functional MRI (fMRI), many studies have been conducted to explore human cognition. The use of fMRI technology is both costly and time-consuming, resulting in relatively small numbers of subjects in treatment conditions—commonly between 10 and 20 subjects per condition. As research has become more sophisticated, there is increasing demand that larger numbers of subjects or patients be included in studies, allowing researchers to investigate both intergroup differences and interindividual differences within groups. As it may be difficult to recruit adequate numbers of volunteers or patients from any given site, and because it is almost impossible to describe population characteristics adequately at a single site, multicenter studies are becoming an increasingly important aspect of neuroimaging research.

Site differences are a particularly important issue in the burgeoning area of cultural neuroscience, the study of cultural differences in neurocognitive processes (1–5). In order to address the critical questions in this subdiscipline, it is typically necessary to collect data from multiple sites that are often located in two or more different countries, and then compare group differences in specific neurocognitive processes. In order to directly compare neural activation patterns across sites (as in cultural studies) or to treat aggregate data from subjects tested at different sites (as in multisite patient studies), it is important to demonstrate that the scanning site is not a significant source of systematic variance in observed neural activation patterns so that one does not falsely conclude that there are differences due to cultural experience that are really due to site.

Given evidence for good reliability of single site results of fMRI for within-subject (for example, Refs. 6,7) and between-subject (for example, Refs. 8,9) activations, multisite studies require reliability of activation detection across sites using different magnets. Early intersite comparisons have largely focused on producing similar thresholded activation maps at different sites for the same stimulus paradigm. Casey et al (10) reported results from four different sites, using different pulse sequences, different fMRI processing software, and 1.5T MR scanners made by two different manufacturers and found reliable patterns of activation across sites. Similarly, Ojemann et al (11) compared fMRI results from two sites with positron emission tomography

[1]Bioengineering Department, University of Illinois at Urbana-Champaign, Urbana, Illinois.
[2]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois.
[3]Department of Psychology, University of Illinois at Urbana-Champaign, Urbana, Illinois.
[4]Cognitive Neuroscience Laboratory, Duke-NUS Graduate Medical School, Singapore.
[5]Center for Brain Health, University of Texas at Dallas, Dallas, Texas.
[6]Radiology, University of Michigan, Ann Arbor, Michigan.
Contract grant sponsor: National Institutes of Health (NIH); Contract grant number: 5R01AG015047-08.
*Address reprint requests to: B.P.S., Bioengineering, 3120 DCL, 1304 W. Springfield Ave., Urbana, IL 61801. E-mail: bsutton@uiuc.edu
Received October 25, 2007; Accepted March 20, 2008.
DOI 10.1002/jmri.21419
Published online in Wiley InterScience (www.interscience.wiley.com).

(PET) results obtained using the same word-stem completion task. Even though the fMRI results originated from two different institutions and two different groups of subjects using MRI hardware produced by different manufacturers with different pulse sequences and analysis techniques, the investigators found highly reproducible areas of activation that agreed well with other findings from PET data. These qualitative findings suggest that fMRI results are fairly robust across a range of conditions.

In light of the promising results from Casey et al (10) and Ojemann et al (11), a number of recent studies have been conducted to provide a concerted effort to examine reproducibility of fMRI responses from different sites. A conservative criterion for the demonstration of multicenter compatibility would be to demonstrate a highly reproducible blood oxygenation level-dependent (BOLD) response rather than simply similar thresholded activation maps, as has been used in previous studies. One of the larger efforts in this direction has been conducted by fBIRN (www.nbirn.net), involving 14 different MR laboratories that use magnets of three different field strengths with three different scanner manufacturers and different pulse sequences (12–17). The initial work has highlighted many factors important to ensure compatibility between different sites. For example, the fBIRN group has shown the importance of simple but regular standardized quality control measures to maintain comparability of scanners using both phantoms (14) and human data (18). In Friedman and Birn (12) and Friedman et al (13), differences in the smoothness and sensitivity of fMRI images were compared among 10 sites. Significant smoothness and sensitivity differences were found that related to imaging sequence, gradient performance, image reconstruction and filtering methods, and field strength. Due to different sensitivities to the BOLD response at different sites, these studies have also motivated an effort to calibrate fMRI responses across sites and individuals using a breath-hold task (15).

In the present study we evaluate the contributions of site, subject, and session to fMRI data variance. Because we are particularly interested in comparing differences in signal patterns in groups tested at two different sites (eg, we are interested in whether East Asians tested in Asia show subtle differences in neural circuitry for processing objects and scenes compared to Westerners tested in the US), it was critically important to demonstrate that observed differences between cultural groups could not be attributed to systematic intersite variability, but rather to systematic differences in neural function between groups. We hypothesized that we would be able to assess intersite reliability if we imaged the same individuals repeatedly at the two sites. Specifically, we hypothesized that we would find that the variability within subjects imaged at two different sites was no greater than the differences within subjects imaged repeatedly at a single site, and less than variability between subjects at a single site. Unlike previous intersite comparisons, we compared sites that had the same imaging hardware, identical pulse sequences, and data analysis strategies using the same subjects at both sites. We will show that within-subject variance across site was small compared to between-subject variance within a single site. This enables quantitative group comparisons across sites using contrasted parameter estimates rather than merely thresholded activation maps.

We examined reproducibility in terms of system stability and noise based on both phantom quality control measures and reproducibility of contrast estimates from motor and visual functional tasks. The motor and visual tasks were performed repeatedly by four subjects and treated in a three-way analysis of variance (ANOVA) to assess main effects and interactions of subject, site, and task.

MATERIALS AND METHODS

The two sites involved in this study, one located in Asia and the other in the US, are henceforth referred to as "site 1" and "site 2." Each site was equipped with a Siemens 3T Allegra MR scanner (Erlangen, Germany) and a USA Instruments (Aurora, OH) headcoil. Protocols for both the phantom quality control (QC) and the human studies were equalized between sites using electronic transfer of protocols.

Visual stimuli were presented using an LCD projector backprojecting onto a screen. The geometry of the screen was matched between sites and the luminance of the projected images was matched using a light meter. Ambient light conditions were matched by running human experiments with scanner room lights off. Auditory stimuli were delivered using an identical headphone system from Resonance Technologies (Northridge, CA).

Phantom Studies

First, we examined system noise using a copper sulfate-doped cylindrical phantom. The phantom is the standard Siemens QA cylindrical phantom. The phantom uses a special holder and markings to ensure proper positioning of the phantom compared to the head coil, so positioning of the phantom between sites was reproducible. The phantom was scanned daily at both MR sites as part of a quality control monitoring routine, with 78 datasets acquired at site 1 and 220 acquired at site 2. The phantom was scanned using an echo planar imaging (EPI) sequence with ramp sampling (36×3 mm thick slices, 0.3 mm slice gap, TR 2 sec, TE 25 msec, flip angle 90°, field of view [FOV] 220×220 mm, 64×64 matrix size, bandwidth of 2894 Hz/pixel). A total of 256 volumes were acquired for a total imaging time of 516 seconds, with 4 seconds of discarded acquisitions. Peak-to-peak (PTP) noise and normalized root-mean-squared (NRMS) noise were used to examine stability in the resulting time courses on a slice-by-slice basis. The PTP and NRMS noise were determined from the time course of the mean over a circular 110 mm area inside the center of phantom. The peak-to-peak noise was calculated by subtracting the minimum value over the mean time course from the maximum and then normalizing by the mean. The NRMS noise was calculated as:

$$NRMS = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(s(t) - \bar{s})^2} \Big/ \bar{s}, \qquad [1]$$

where $s(t)$ is the value of the mean time course at volume t, $\bar{s}$ is the mean of the mean time course, and T is the total number of volumes acquired.

The mean, NRMS, and PTP noise were compared between sites to determine if there was a system noise difference between the two sites. This system noise represents only one component of the noise in a functional imaging data, with physiological noise potentially dominating system noise. The phantom analysis can serve to indicate when system maintenance is required by examining the daily noise properties versus the history or by comparing the system noise to residual mean error after fitting functional imaging data.

Reproducibility of Contrast Estimates as a Function of Subject and Site

Two tasks were presented to four subjects scanned at both sites—a motor task and a visual task. The functional acquisition was an EPI sequence with ramp sampling (32 × 4 mm thick slices, 0.4 mm slice gap, TR 2 sec, TE 25 msec, flip angle 80°, FOV 220 × 220 mm, 64 × 64 matrix size, bandwidth 2894 Hz/pixel). Four subjects (males, ages 24, 26, 27, and 28; mean 26.25 years) were repeatedly scanned at both sites on the two tasks (same four subjects at both sites), with the task design modeled after McGonigle et al (6). All subjects were right-handed and visual acuity was corrected with MR-compatible lenses to 20/30 using a Snellen Chart. For each task there were ≈30 scan sessions per subject (15 at each site) spread over 3 days at each site, with a total of 230 scan sessions across all tasks and subjects. The first task was a motor task: button pressing paced by auditory tones (1 Hz). Subjects performed alternating 20-second blocks of button pressing and rest with a total of six button-pressing blocks and seven rest blocks in each session; 3 seconds of visual instructions (indicating whether to press or rest) preceded each block. A single finger on each hand was used for pressing a button on a response box (Rowland Institute USB fMRI response boxes, Cambridge, MA). The second task was a visual paradigm with 20-second blocks alternating between a fixation cross and an 8 Hz reversing checkerboard. Three rest and three checkerboard blocks were presented to the subject per run.

During each scan session the subjects underwent a localizer scan in addition to performing the motor and visual tasks. A T2-weighted turbo-spin-echo (TSE) high-resolution anatomical scan with the same slice prescription as the EPI acquisition was used for image coregistration. The subjects were removed from the scanner between each scan session. The 15 sessions at each site were spread out over 3 days of scanning. During one session a magnetization-prepared rapid acquisition of gradient echo (MPRAGE) sequence was used to acquire a high-resolution 3D structural scan used for normalization. The scan acquired whole brain with isotropic 0.8 mm resolution and an inversion preparation time of 1100 msec.

Functional analysis was carried out using FEAT (FMRI Expert Analysis Tool) v. 5.4, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The following preprocessing steps were applied at the first-level analysis: slice-timing correction using Fourier-space time-series phase-shifting; motion correction using MCFLIRT (19); nonbrain removal using BET (20); spatial smoothing using a Gaussian kernel of full-width at half-maximum (FWHM) 8 mm; mean-based intensity normalization of all volumes by the same factor; high-pass temporal filtering (Gaussian-weighted LSF straight line fitting, with sigma = 50.0 sec). Time-series statistical analysis was carried out using FILM with local autocorrelation correction (21). Registration proceeded in three steps using FLIRT (19,22). First, the EPI images were registered to the TSE scans, which were then registered to the MPRAGE anatomical image, and finally to the standard (MNI) image. Higher-level analysis was carried out using FLAME (FMRIB's Local Analysis of Mixed Effects) (23,24). A fixed effects three-way ANOVA was performed in MatLab (MathWorks, Natick, MA) on all 230 functional runs from the study (4 subjects × 2 tasks × 2 sites with runs [13–15 per subjects] as a random variable) to assess the effect of subject, site, and task on the functional results.

Besides qualitatively examining the maps of significant voxels in the main effects and interactions resulting from the ANOVA analysis (thresholded at $P < 0.001$ uncorrected), two other analyses of the functional data were performed. We examined the spatial extent of activation in single subjects across the two sites and we analyzed the percent signal change over a visual and two motor regions-of-interest (ROIs) for each subject at each site. Additionally, the effect size from the ANOVA was examined to determine the proportion of variance due to site as compared to intersubject variance.

RESULTS

Signal Stability in Phantoms and Volunteers

The resulting mean of the time series and NRMS noise values from several months of quality control scans of the cylindrical phantom at both sites are shown in Fig. 1 as strip plots. The PTP noise looked similar to the NRMS noise and is not shown. The phantom result images show the mean and NRMS noise for each slice down the columns and for each day across the rows. Note that at each site there were a few days or clusters of days that had noise values that were significantly higher than surrounding days. This suggests the importance of daily quality control. On days when the noise was unacceptably high the MRI service engineer was notified that maintenance was required on the system and all scanning sessions were cancelled for that day.

Although the means of the time series appear equivalent across sites, Fig. 1 shows obvious differences between the noise characteristics. In fact, grouping all 36 slices into a t-test between the two sites resulted in significant differences for mean, NRMS noise, and PTP noise (all at $P < 0.01$). Figure 2a plots the results of NRMS noise as a percent of the mean signal for slice 18 of 36. Note that more phantom scans were performed at site 2 than site 1. The differences in noise may result from actual system noise, environmental differences between scanning sites, or a difference in the number of

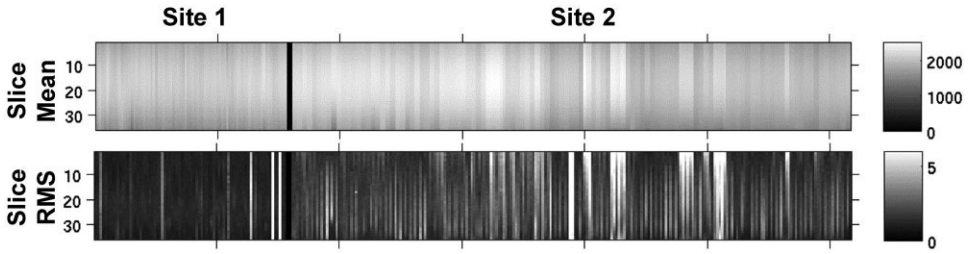

Figure 1. Phantom quality control runs at both sites, showing the mean and noise properties across all 36 slices of the QA acquisitions. Top row is mean and bottom row is root mean square noise of the time series of the averaged ROI.

items passed through the patch panel that may leak small amounts of noise.

At first glance, these results might seem to suggest an inconsistency between the systems. However, the system noise must be considered relative to the physiological noise present in a functional run. The physiological noise may dominate the relevant noise properties of a functional run, rendering slight differences in system noise as relatively inconsequential. In order to examine the relationship between this system noise and the noise in the functional runs, we examined fMRI noise at two levels: 1) looking at the supraventricular white matter in unprocessed functional data, and 2) examining the residual error after the first-level functional analysis in the 230 runs from our four volunteers.

Signal noise was analyzed by calculating the NRMS noise from the supraventricular white matter in the raw functional images. Data in this region are not affected by function, but reflect the stability in the baseline signal. A conservative mask was made of the supraventricular white matter consisting of 2192 voxels in the high-resolution standard space (2 mm isotropic resolution). After motion-correcting the raw functional time series relative to the middle volume, the white matter mask was transformed into the subject's local space by using the FSL transformations determined during the first level analysis. For each timepoint in the functional series the mean over the white matter mask was calculated. The time series of the mean was transformed into NRMS according to Eq. [1]. The NRMS as a percentage of the mean is plotted in Fig. 2b.

Additionally, for each functional run the residual error (σ^2) was computed for all voxels that showed significant activation ($P < 0.001$ uncorrected) to the task in a single run at the first-level analysis. This residual error map was formed into an NRMS measure by averaging the σ^2 across voxels, taking the square root, and normalizing by the mean of the functional imaging time course. The plot of NRMS error from the functional scans is plotted in Fig. 2c.

While it is tempting to quantitatively compare the NRMS measure from the human scans to those in Fig. 2a from the phantom scans, note, however, that the slice thickness is different between the phantom and human acquisitions. Instead, it is better to compare the NRMS error in Fig. 2b,c between sites. Note that no significant differences existed in this error measure between sites ($P = 0.38$ for both), while differences between subjects were observable. This indicated that the differences between system noise were insignificant,

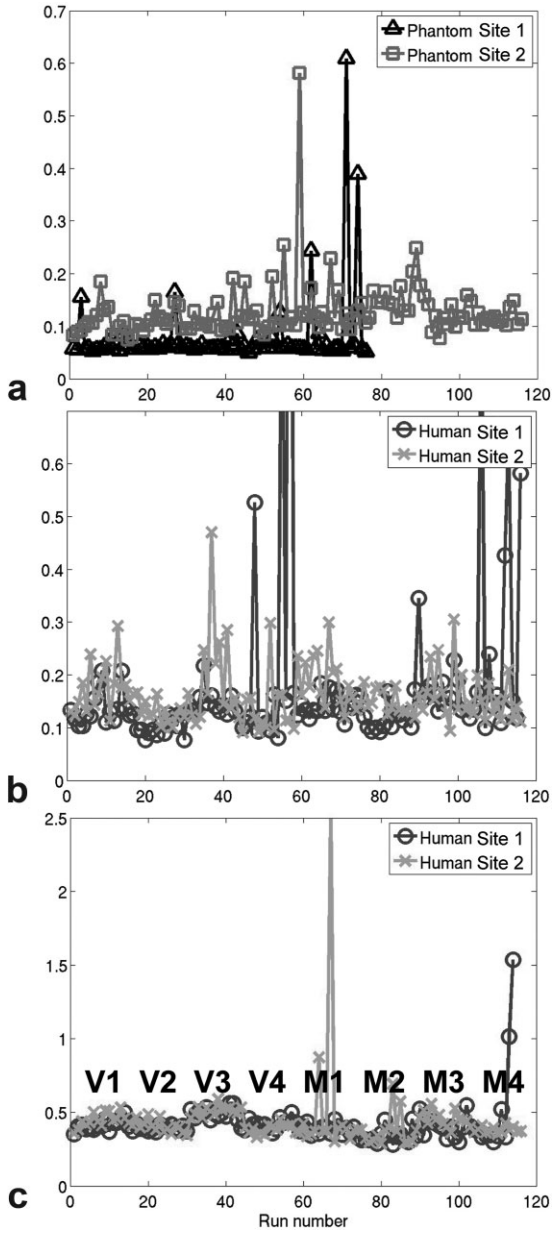

Figure 2. Normalized root mean square error (or noise) as a percentage of the mean for **(a)** phantom quality control runs, **(b)** supraventricular white matter, and **(c)** residual mean error in the motor and visual functional tasks. Note that V1 indicates subject 1, visual task and M2 indicates subject 2, motor task. Subject/task labels are valid for (b) and (c). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

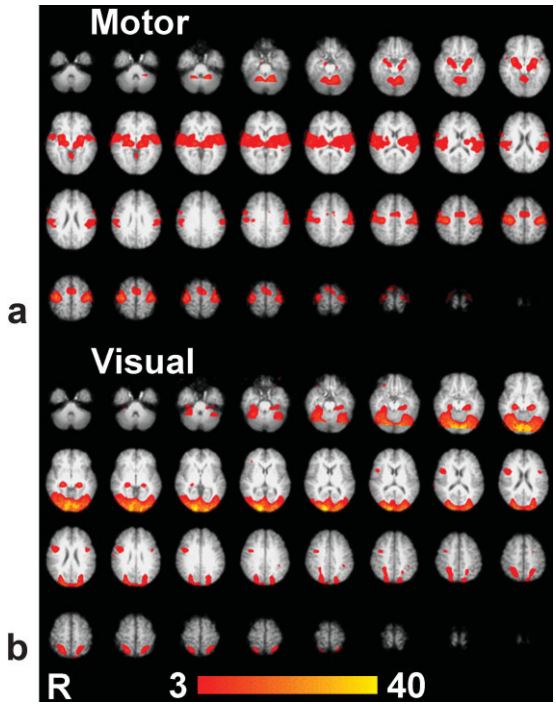

Figure 3. Thresholded z-scores for all subjects for each task ($P < 0.001$ uncorrected).

dominated by other sources of noise, and did not affect the ability to quantitatively compare the functional results across sites.

Statistical Maps

A three-way ANOVA analysis was performed in Matlab on the 230 functional runs with main effects of subject, site, and task and their interactions being investigated. For reference, thresholded ($P < 0.001$ uncorrected) z-score maps of the mean activations for each task over all subjects and both sites are shown in Fig. 3. The robust activations resulted from the two reproducible tasks used in this analysis and the 115 functional runs used per map. The thresholded ($P < 0.001$ uncorrected) ANOVA results for the main effects of subject, site, and task are shown in Fig. 4. There were extensive regions showing significant main effects of subject and task, but there were very few voxels with main effect of site surviving $P < 0.001$ uncorrected. The main effect of task occurred as expected, due to the significantly different brain areas activated by the motor versus visual tasks. These large task differences were included in the ANOVA analysis for completeness and to remove these effects as a source of variance. The site effects were minimal—there were several regions that indicated differences: regions at the most inferior and posterior parts of the brain. These effects, however, were quite modest and sparse when compared to the subject differences, which were much larger, as shown in Table 1, which is described later. Slight differences in placement of the subject's head could have resulted in these differences at the lower periphery of the FOV due to shimming differences, RF coil sensitivity, or gradient nonlinearity effects. The scanner's automatic gradient

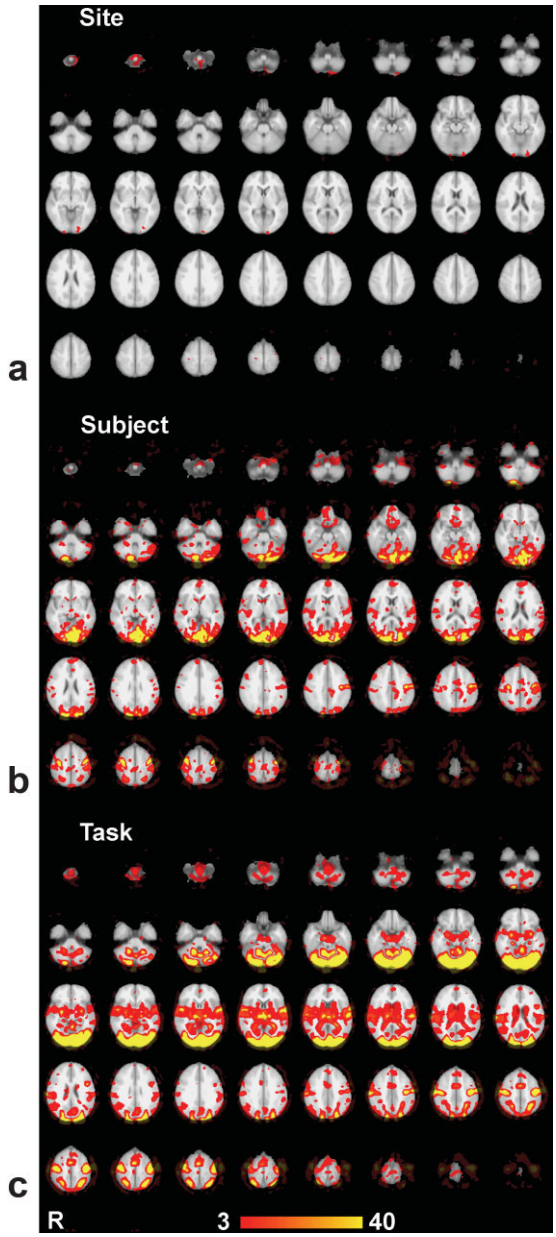

Figure 4. Main effects of the three-way ANOVA, thresholded with $P < 0.001$ (uncorrected). Results are displayed as $-\log_{10}(p)$ shown from $P =$ 10e-3 to 10e-40.

shimming routine was employed to shim over the prescribed slices; however, differences in head placement could affect the shimming convergence. Gradient nonlinearity effects can be addressed by warping algorithms and this has shown promising results for 3D morphometry studies (25). Validation of these methods

Table 1
Effect Size Analysis From the ANOVA Results for Site and Subject

	Site		Subject	
	Mean ωG2	Max ωG2	Mean ωG2	Max ωG2
Visual	0.0112	0.0356	0.0810	0.2342
Motor left	0.0120	0.0248	0.0966	0.2574
Motor right	0.0085	0.0152	0.0880	0.2501

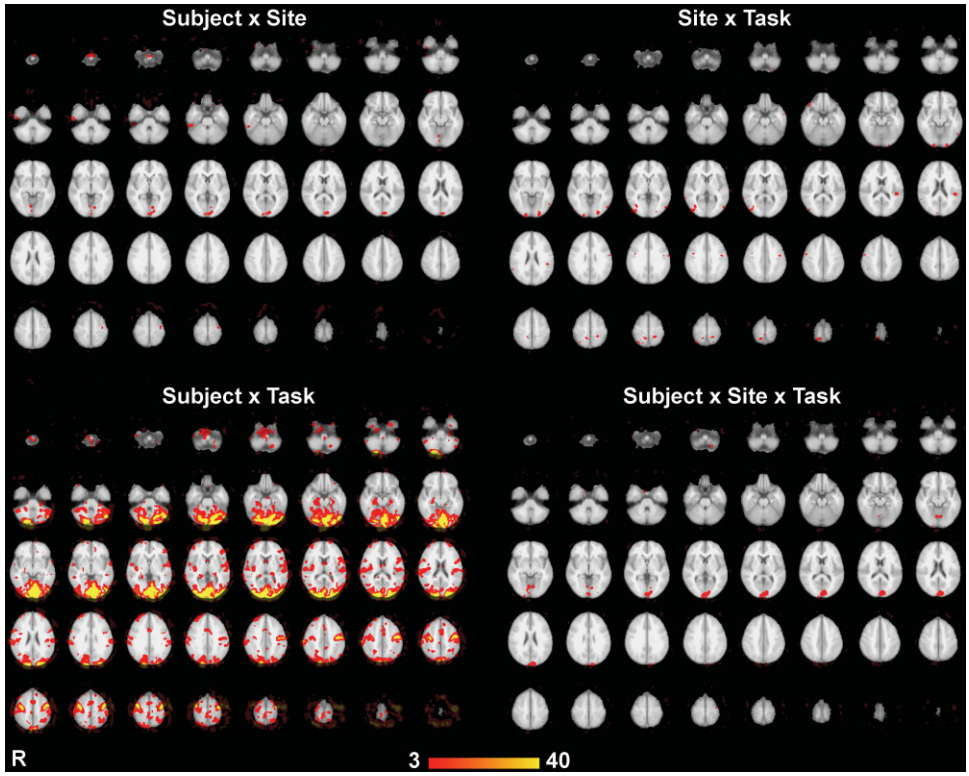

Figure 5. Interaction effects from the ANOVA, thresholded with $P < 0.001$ (uncorrected). Results are displayed as $-\log_{10}(p)$ shown from $P =$ 10e-3 to 10e-40.

for functional imaging applications is still lacking; therefore, extra care in placement of the subject's head will be exercised in future studies.

The interaction effects are shown in Fig. 5. The most striking effect is that there was a pronounced Subject × Task interaction, but that interactions involving site occurred in relatively few brain areas. Specifically, we found a significant three-way interaction among site, subject, and task in the medial occipital region (Fig. 5, top left panel). Examining the two-way interactions suggested that this might stem from the difference in activation across subjects in these regions being significantly different across sites (Fig. 5, top right). However, note that there was a greater interaction effect between subject and task (Fig. 5, lower left), suggesting that the differences across site did not affect our ability to detect interindividual or task differences when data from both sites were collapsed together.

In order to assess functional sensitivity at both sites we compared the spatial extent of activation across sites. Using *cluster* in FSL we computed the number of voxels in the largest cluster using a threshold of $P <$ 0.001, uncorrected, for each subject and each task. A cluster includes all contiguous voxels that are below the threshold of 0.001. Then we performed two-sample *t*-tests between sites for the cluster sizes for each of the approximately 15 runs. The *P*-value for the *t*-test is given in Table 2 for each subject and task. Only one of the eight pairings yielded statistically significant differences, suggesting that the spatial extent is well matched between sites. As further evidence, Fig. 6 shows a scatterplot of the pairings of cluster sizes between site 2 vs. site 1. The scatterplot shows pairings of run 1 at site 1 with run 1 at site 2, etc. This scatterplot indicates that there is good agreement between sites in the number of voxels in the largest cluster.

An ROI analysis was performed on percent signal change data from three functionally defined areas, one in the visual cortex, one in the right motor area, and one in the left motor area. The ROIs were defined by thresholding the group mean activation z-scores in the visual and motor tasks at a level near the peak, then eroding and dilating the mask such that only one contiguous region of voxels was selected around the peak of interest. The number of voxels in the visual ROI was 63, in the left motor 102 voxels, and in the right motor 54 voxels. The ROIs were used to examine percent signal change at both sites. A pairing of run 1 at site 1 with run 1 at site 2, and so on, resulted in the scatterplots in Fig.

Table 2
Results From Two-Sample t-Test for Comparison of Maximum Cluster Size Between Two Sites

	Task	t-Statistic	P-value
Subject 1	Motor	0.398	0.6937
	Visual	1.556	0.1309
Subject 2	Motor	0.274	0.7863
	Visual	1.526	0.1381
Subject 3	Motor	0.513	0.6124
	Visual	2.572	0.0164*
Subject 4	Motor	0.049	0.9612
	Visual	0.338	0.7384

*Significance at the 0.05 level.

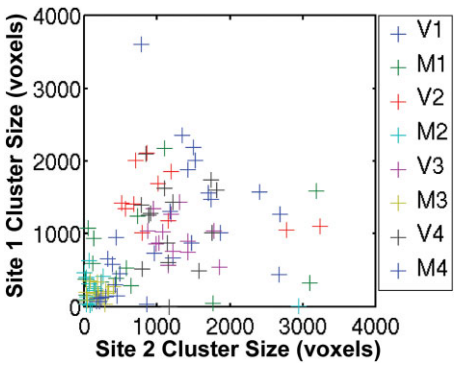

Figure 6. Scatterplot of number of voxels in largest cluster at site 1 vs. site 2. Data labels are defined in Fig. 2.

7. As can be seen in these figures, the percent signal changes were correlated between sites and intersubject differences were readily observable beyond any site differences. This was especially true in the visual ROI, as the activations were much stronger than in the motor areas. Also notice that the left motor cortex activation shows higher consistency than the right motor cortex, likely reflecting the right-handedness of the subjects. The average percent signal change inside the visual ROI for every visual task run was compared between sites with a two-sample *t*-test to assess if the means were equal. Likewise, the motor ROIs were used to calculate average percent signal change from motor runs and assessed for equal means via a two-sample *t*-test. The result (dof = 112 for all ROIs) for the Visual ROI was $P = 0.89$, Left Motor ROI $P = 0.16$, Right Motor ROI $P = 0.38$. This provides further evidence suggesting that the site effect is smaller than subject-related variability in ROI analyses.

Finally, the ROIs above were used to examine generalized effect size of the site and subject components from the ANOVA. We used the generalized ω_G^2 from Olejnik and Algina (26) to provide a measure of effect size that should be robust to the specific effects (including task) that were included in the ANOVA analysis. The generalized ω_G^2 describes the proportion of the variance explained by each effect. We treated site as a manipulated factor and subject as a measured factor and used the formula provided in table 2 of Olejnik and Algina (26). The results in Table 1 show that the site accounted for around 1% of the variance, whereas subject accounted for 8%–10%. This suggests that the effect of site is much smaller than the intersubject variance.

DISCUSSION

A methodology is currently being developed to take advantage of the collective resources available across multicenter collaborations. Currently these collaborations involve different imaging hardware platforms, different sequences, and different data analysis strategies. Quantitative comparison of results from these collaborations will rely on either signal processing to make signals comparable from different sites or on calibration signals obtainable during a normal functional evaluation (15). In the current study we showed that quantitative comparisons of functional data can be made without calibration steps if two sites are matched on imaging hardware, sequences, analysis methods, scanner performance, and quality control.

Our results examine comparability of fMRI data generated at two sites in terms of phantom noise and stability measurements as well as in fMRI experiments. The phantom data show that it is important to monitor system noise to ensure that it is low. With careful quality control measures, however, the impact of small differences in system noise will not have an impact on the noise level in functional imaging data, which is largely dominated by physiological noise. The three-way ANOVA analysis from repeated scanning of four subjects on two tasks indicated a much greater main effect of subject than site. However, there were some significant voxels showing site differences, especially inferior regions and cerebellum. This could be due to systematic head placement differences due to different subject positioning pads, an effect that will be controlled in future studies. The main effect and interaction effects that include site show very small numbers of significant voxels ($P < 0.001$ uncorrected) with low mean z-scores compared to other effects. By comparing generalized effect sizes, we see that there is a nearly 10 times greater effect of subject than site, rendering the platform reliable for comparisons using group studies.

Although we matched some parameters of the task performance, even more careful control in even these simple tasks could provide more information about the contribution of intra- and intersubject variability to the

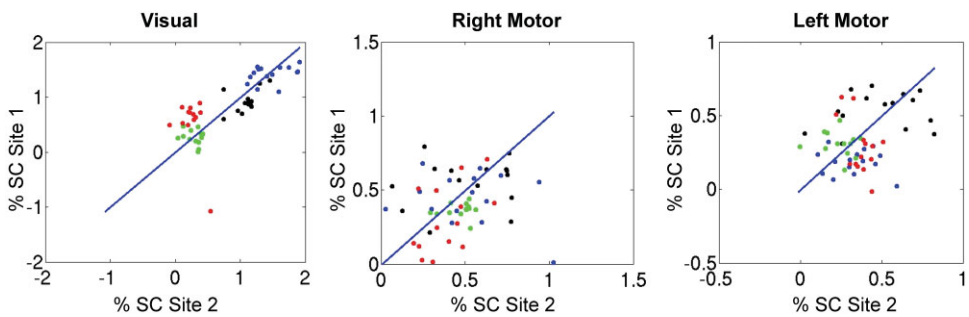

Figure 7. Scatterplot of percent signal change (% SC) in ROIs for the visual, right motor, and left motor activations. The scatterplots are mean percent signal change at site 1 vs. site 2. Individual subject scatterplots are evident by color-coding the subjects as follows: Subject 1 (Black), Subject 2 (Blue), Subject 3 (Green), Subject 4 (Red).

noise in functional imaging data. We matched the luminance and geometry of the projection for the visual stimulus and, as shown in Fig. 6, the visual task shows higher reproducibility than the motor task. Monitoring of visual attention with an eye tracker and using this information as a covariate in the functional analysis could further reduce the subject variability. For the motor task we used identical button boxes and pacing of the motor task; however, perhaps even more variance would have been explained if we had controlled for finger pressure, extension angle, and acceleration.

This reproducibility comparison using the same imaging hardware and analysis strategy is important not only for the intersite cultural comparison being investigated here, but also for long-term longitudinal (27) and intervention fMRI studies (28). In such studies it is assumed that the state of the system, with the same hardware, has not changed over time. Hardware and software upgrades over the life cycle of a system may require calibration measures similar to Thomason et al (15) to account for changes in sensitivity of the system. However, an assessment of the system noise contribution, similar to that described in this work, could serve as an important baseline acceptance test. Assessing quantitative contributions of system noise for longitudinal studies may require more careful control of the tasks to reduce performance-induced variance and fatigue effects. Such controls may include using an eye tracker; controlling the angle, force, and velocity of the motor tasks; and using behavioral measures as correlates in the functional analysis. Further study is required to examine comparability of "matched" systems over time.

In conclusion, we have shown that reliable, quantitative imaging results can be obtained in multicenter studies using the same imaging platform, sequences, and data analysis methods. With attention to both phantom and human functional signal quality control, multicenter collaborations of matched sites can exist and perform quantitative comparisons of functional signals where site is confounded with group. This analysis lays the foundation for future studies that utilize these scanners to perform cross-cultural quantitative statistical analyses where the populations to be studied are separated by large geographical distances.

REFERENCES

1. Hedden T, Park DC, Nisbett R, Ji LJ, Jing Q, Jiao S. Cultural variation in verbal versus spatial neuropsychological function across the life span. Neuropsychology 2002;16:65–73.
2. Nisbett RE, Miyamoto Y. The influence of culture: holistic versus analytic perception. Trends Cogn Sci 2005;9:467–473.
3. Gutchess A, Welsh R, Boduroglu A, Park DC. Cultural differences in neural function associated with object processing. Cogn Affect Behav Neurosci 2006;6:102–109.
4. Park DC, Gutchess AH. The cognitive neuroscience of aging and culture. Curr Directions Psychol Sci 2006;15:105–108.
5. Goh JO, Chee MW, Tan JC, et al. Age and culture modulate object processing and object-scene binding in the ventral visual area. Cogn Affect Behav Neurosci 2007;7:44–52.
6. McGonigle DJ, Howseman AM, Athwal BS, Frackowiak RS, Holmes AP. Variability in fMRI: an examination of intersession differences. Neuroimage 2000;11:708–734.
7. Smith SM, Beckmann CF, Ramnani N, et al. Variability in fMRI: a re-examination of inter-session differences. Hum Brain Mapp 2005;24:248–257.
8. Cohen MS, DuBois RM. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J Magn Reson Imaging 1999;10:33–40.
9. Chee MW, Lee HL, Soon CS, Westphal C, Venkatraman V. Reproducibility of the word frequency effect: comparison of signal change and voxel counting. Neuroimage 2003;18:468–482.
10. Casey BJ, Cohen JD, O'Craven K, et al. Reproducibility of fMRI results across four institutions using a spatial working memory task. Neuroimage 1998;8:249–261.
11. Ojemann JG, Buckner RL, Akbudak E, et al. Functional MRI studies of word-stem completion: reliability across laboratories and comparison to blood flow imaging with PET. Hum Brain Mapp 1998;6:203–215.
12. Friedman L, Birn F. An index of scanner/site differences in fMRI sensitivity: method and implications. In: Proc ISMRM 2004;11:489.
13. Friedman L, Magnotta VA, Posse S, Birn F. Scanner differences in the smoothness of fMRI images: implications for multi-center studies. In: Proc ISMRM 2004;11:1074.
14. Glover GH, Foland L, Birn F. Scanner quality assurance for longitudinal or multicenter fMRI studies. In: Proc ISMRM 2004;11:992.
15. Thomason ME, Foland L, Glover GH. Calibration of BOLD fMRI using breath holding reduces group variance during a cognitive task. Hum Brain Mapp 2007;28:59–68.
16. Zou KH, Greve DN, Wang M, et al. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by biomedical informatics research network. Radiology 2005;237:781–789.
17. Zou KH, Greve DN, Wang M, et al. A prospective multi-institutional study of the reproducibility of fMRI: a preliminary report from the biomedical informatics research network. Lecture Notes in Comput Sci 2004;3217:769–776.
18. Stocker T, Schneider F, Klein M, et al. Automated quality assurance routines for fMRI data applied to a multicenter study. Hum Brain Mapp 2005;25:237–246.
19. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002;17:825–841.
20. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17:143–155.
21. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 2001;14:1370–1386.
22. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5:143–156.
23. Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in FMRI. Neuroimage 2003;20:1052–1063.
24. Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for FMRI group analysis using bayesian inference. Neuroimage 2004;21:1732–1747.
25. Jovicich JS, Czanner, Greve D, et al. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 2006;30:436–443.
26. Olejnik S, Algina J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 2003;8:434–447.
27. Gunter JL, Shiung MM, Manduca A, Jack CR Jr. Methodological considerations for measuring rates of brain atrophy. J Magn Reson Imaging 2003;18:16–24.
28. Aron AR, Gluck MA, Poldrack RA. Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage 2006;29:1000–1006.