NeuroImage: Clinical 5 (2014) 309–321

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

Intersession reliability of fMRI activation for heat pain and motor tasks

Raimi L. Quiton[a,b,c,*], Michael L. Keaser[c], Jiachen Zhuo[d], Rao P. Gullapalli[d], Joel D. Greenspan[c]

[a]Department of Psychology, University of Maryland, Baltimore County, MD, USA
[b]Department of Anatomy and Neurobiology, School of Medicine, University of Maryland, Baltimore, MD, USA
[c]Department of Pain and Neural Sciences, School of Dentistry, and UM Center to Advance Chronic Pain Research, University of Maryland, Baltimore Country, MD, USA
[d]Department of Diagnostic Radiology & Nuclear Medicine & Magnetic Resonance Research Center, School of Medicine, University of Maryland, Baltimore Country, MD, USA

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 22 June 2014
Accepted 17 July 2014
Available online 22 July 2014

Keywords:
Intraclass correlation coefficient
Reliability coefficient
Reproducibility
Repeatability
Anterior insula
Cingulate cortex

ABSTRACT

As the practice of conducting longitudinal fMRI studies to assess mechanisms of pain-reducing interventions becomes more common, there is a great need to assess the test–retest reliability of the pain-related BOLD fMRI signal across repeated sessions. This study quantitatively evaluated the reliability of heat pain-related BOLD fMRI brain responses in healthy volunteers across 3 sessions conducted on separate days using two measures: (1) intraclass correlation coefficients (ICC) calculated based on signal amplitude and (2) spatial overlap. The ICC analysis of pain-related BOLD fMRI responses showed fair-to-moderate intersession reliability in brain areas regarded as part of the cortical pain network. Areas with the highest intersession reliability based on the ICC analysis included the anterior midcingulate cortex, anterior insula, and second somatosensory cortex. Areas with the lowest intersession reliability based on the ICC analysis also showed low spatial reliability; these regions included pregenual anterior cingulate cortex, primary somatosensory cortex, and posterior insula. Thus, this study found regional differences in pain-related BOLD fMRI response reliability, which may provide useful information to guide longitudinal pain studies. A simple motor task (finger-thumb opposition) was performed by the same subjects in the same sessions as the painful heat stimuli were delivered. Intersession reliability of fMRI activation in cortical motor areas was comparable to previously published findings for both spatial overlap and ICC measures, providing support for the validity of the analytical approach used to assess intersession reliability of pain-related fMRI activation. A secondary finding of this study is that the use of standard ICC alone as a measure of reliability may not be sufficient, as the underlying variance structure of an fMRI dataset can result in inappropriately high ICC values; a method to eliminate these false positive results was used in this study and is recommended for future studies of test–retest reliability.

1. Introduction

Functional magnetic resonance imaging (fMRI) studies in chronic pain patients have the potential to provide valuable information about the neural mechanisms and efficacy of analgesic therapies, including drug treatments, acupuncture, brain stimulation, distraction tasks, mindfulness meditation, and cognitive-behavioral interventions. Furthermore, fMRI studies in healthy individuals have provided insights into neural mechanisms of pain modulation, such as the placebo effect and conditioned pain modulation. Such studies rely on the assumption that brain responses to pain are consistent in sessions conducted on separate days before, during, and after therapeutic interventions. Despite this common practice, only one study specifically addressing the intersession reliability of pain-related fMRI activation has been published. Taylor and Davis (2009) examined the spatial reliability of fMRI activation associated with painful mechanical stimulation of the hand in 6 subjects across four biweekly sessions, finding high across-session spatial repeatability in second somatosensory cortex (S2), but lower and more variable spatial repeatability in primary somatosensory cortex (S1) and thalamus (Taylor and Davis, 2009); other areas that are part of the cortical network classically activated by pain (as reviewed in Duerden and Albanese, 2013) were not examined in the study. Furthermore, studies examining across-session reliability of the amplitude (percent signal change) of pain-related fMRI responses have not yet been published. Gaining a better understanding of the stability of repeated, intersession measures of responses to pain (both spatial extent and BOLD signal amplitude) in the entire cortical pain network will enhance the ability to interpret data collected in studies of pain-reducing manipulations. Thus, characterization of test–retest reliability of pain-related fMRI activation is a critically important issue to address.

High test-retest reliability of fMRI responses across two or more sessions has been reported for a wide variety of tasks, including motor, auditory detection, language, learning, and memory (Atri et al., 2011; Bennett and Miller, 2010; Cacares et al., 2009; Chen et al., 2007; Fliessbach et al., 2010; Freyer et al., 2009; Gorgolewski et al.,

* Corresponding author.
E-mail address: rquiton1@umbc.edu (R.L. Quiton).

http://dx.doi.org/10.1016/j.nicl.2014.07.005

2013; Gountouna et al., 2010; Havel et al., 2006; Kiehl and Liddle, 2003; Maitra et al., 2002; McGregor et al., 2012; Yoo et al., 2005). However, poor test–retest reliability of fMRI responses has been found for other tasks such as mouth movements (Havel et al., 2006), reward (Fliessbach et al., 2010), and spatial attention (Gorgolewski et al., 2013). While these results may reflect differences in reliability across tasks, they may also reflect differences in statistical approaches used to assess reliability. Approaches have included measuring spatial extent or spatial overlap of significant task-related activation at the individual or group level, performing voxelwise group-level contrasts of the blood oxygenation level-dependent (BOLD) response, and calculating voxel-wise intraclass correlation coefficients (ICC) based on the amplitude of the BOLD response.

The first objective of this study was to characterize the intersession reliability of pain-related fMRI activation elicited by painful contact heat stimuli in healthy volunteers, considering both spatial extent and amplitude measures. To address the possibility that reliability of pain-related fMRI responses may vary across brain regions, the study separately evaluated reliability of responses in cortical areas that are part of the network classically activated by pain (as reviewed in Duerden and Albanese, 2013), including: S1, S2, pregenual anterior cingulate cortex (pACC), anterior midcingulate cortex (aMCC), insular cortex (INS, distinguishing anterior and posterior), supplementary motor area (SMA), and several frontal lobe regions; the thalamus was also examined.

Stimulus paradigms using painful heat vary across studies, with some protocols using fixed temperatures for all subjects (resulting in highly variable perceived pain intensity reports across the group) and others using subject-specific temperatures that produce consistent perceived pain intensities across the group. A recent study reported no difference in pain-related fMRI activation produced by fixed temperature stimuli and individually-determined contact heat pain stimuli (van den Bosch et al., 2013); however, the question of whether these different stimulus paradigms produce equally reliable results across sessions remains unanswered. A second objective of this study was to compare the reliability of pain-related fMRI responses for fixed temperature stimuli with that of subject-specific temperatures to address whether one or the other approach provides for more reproducible results.

The subjects in this study also performed a simple motor task (finger-thumb opposition) in the same sessions as they experienced painful heat. To evaluate the appropriateness and validity of the analytical approach used to assess intersession reliability of pain-related fMRI responses, this study used the same approach to assess the across-session reliability of motor-related fMRI activation and compared the results with those previously published.

2. Methods

2.1. Subjects

Fourteen subjects (mean age 44.3 years, SD 19, range 22–75; 7 male) participated in the study. All subjects were healthy, with no major medical, neurological, or chronic pain disorders. Young female subjects were tested during the follicular phase of their menstrual cycle (days 3 to 10) to reduce variance potentially related to effects of gonadal hormone fluctuations on pain perception. All postmenopausal women ($n = 4$) were not using hormone replacement therapy. Informed consent was obtained from all subjects prior to experimentation. The protocol for this study was approved by the University of Maryland Institutional Review Board for the Protection of Human Subjects.

2.2. Stimulation

Painful heat stimuli were delivered to the left dorsal forearm of each subject using an MR-compatible Peltier thermal probe with a 2.6 cm^2 contact surface (TSA-II, Medoc Ltd., Israel). The probe was held in place during testing with a Velcro strap. Two temperatures of painful heat were delivered to each subject: (1) 48 °C and (2) a subject-specific temperature perceived as moderately painful, which was defined as the temperature the subject rated as 50 on a 100-point computerized visual analog scale (VAS) for pain intensity. These stimuli will be referred to as 48 °C and 50VAS throughout this manuscript. Temperatures that evoked the perception of moderate pain (50VAS) ranged from 47.5 to 50.0 °C (mean 49.0 °C, SD 1.0).

2.3. Experimental Protocol

2.3.1. Training session

Subjects participated in a training session in a laboratory room dedicated to psychophysical assessments at least one day prior to scanning. During the training session, subjects were first presented with an ascending series of thermal stimuli ranging from 42 to 50 °C. Each temperature was presented for 15 s (including ramp up and down time at a rate of 2.7 °C/s), followed by a 30-s interstimulus period of nonpainful warmth (37 °C). Subjects were then presented with a series of heat stimuli expected to be in the painful range (46–50 °C), with temperatures presented twice each in a randomized order. After each stimulus, subjects used an MR-compatible trackball (Fellowes, *http://www.fellowes.com*) to rate peak pain intensity on a computerized VAS, which consisted of a vertical scale labeled "no pain" at the bottom and "most intense pain imaginable" at the top (DAPSYS, Brian Turnquist, Johns Hopkins University, *http://www.dapsys.net*). VAS ratings were converted to numerical values ranging from 0 to 100. Individual subject ratings for the range of temperatures were used to interpolate the subject-specific temperature that evoked a perception of moderate pain (50VAS) and to confirm that the 48 °C stimulus was perceived as painful.

2.3.2. Scanning sessions

Each subject participated in three scanning sessions conducted on separate days, with the mean interval between sessions 15 days (SD 18). The high variance in the between-session interval was mainly attributable to two women who were scanned across multiple months to ensure testing was conducted during the follicular phase of the menstrual cycle. Most scans occurred between 4 and 9 pm and lasted about 90 minutes; each of the 3 sessions for an individual subject began at about the same time each day to reduce circadian variability in perception and hormone levels. During each scanning session, information about functional brain responses was collected using BOLD fMRI and information about brain anatomy was collected using MR imaging (details below). The fMRI portion of the session consisted of two scans in which painful heat stimuli were delivered, separated by a 30-minute interval. The painful heat stimulus protocol for each fMRI scan consisted of delivering the two temperatures (48 °C and the subject-specific temperature perceived as moderately painful) six times each in a randomized order. Each temperature was presented for 15 s (including ramp up and down time at a rate of 2.7 °C/s), followed by a 30-s interstimulus period of nonpainful warmth (37 °C). After each stimulus, the computerized VAS was presented to the subject through MR Vision 2000 goggles (Resonance Technologies, Van Nuys, CA) and the subject rated peak pain intensity using the MR-compatible trackball. The duration of each stimulus cycle was 45 s: painful heat application (15 s), VAS rating task (15 s, or less if the subject responded more rapidly), and rest period (15 s, or more if the subject completed the rating task in less than 15 s).

In the 30-minute interval between pain fMRI scans, subjects rested quietly for approximately 10 minutes, performed a simple motor task (right hand finger-thumb opposition at approximately 1 Hz) in a block design (24 s opposition alternating with 24 s rest) for 6 minutes and 54 s while fMRI data were acquired, then rested quietly for the remainder of the interval.

2.3.2.1. Image acquisition. Images were collected using a 1.5 Tesla Phillips Eclipse scanner (Phillips Healthcare, Cleveland, OH). Functional MR

images were acquired using a single-shot echo planar imaging T2*-weighted sequence with an echo time (TE) of 35 ms and flip angle of 90°. Acquired image resolution was 3.2 × 3.2 mm over a 24-cm field-of-view (FOV). The images were zero padded to 128 × 128 pixels to provide a resolution of 1.875 × 1.875 mm. The repetition time (TR) of 3 s provided full brain coverage using 24 axial slices (6 mm thick with no gaps between slices) prescribed parallel to the anterior-posterior commisural plane. High-resolution T1-weighted volumetric scans (4.5 ms TE, 29 ms TR, 110 slices, slice thickness 1.5 mm, 0.938 × 0.938 mm in-plane resolution, 24-cm FOV) were acquired in the same plane as the functional images for anatomical detail.

2.3.2.2. Image processing and analysis

2.3.2.2.1. Preprocessing. Image processing was performed using the software package Analysis of Functional Neuroimages (AFNI; Cox, 1996). The first four volumes of each functional scan were discarded to exclude images acquired prior to stabilization of the magnetic resonance signal. The remaining volumes were corrected for slice timing differences. Data from the 2 functional pain scans conducted within each session were concatenated for analysis. Functional images were motion-corrected by spatially registering the volumes from all functional scans to the first remaining volume (AFNI routine 3dVolreg). To minimize effects from possible spike-related artifacts, signals greater than 2.5 SDs of the overall BOLD signal were reduced (AFNI routine 3dDespike). Time series were temporally smoothed to reduce high frequency noise using a moving 3-point weighted (0.15-0.70-0.15) average. Images were spatially smoothed to increase the signal-to-noise ratio using a 5-mm full width half-maximum Gaussian kernel. Trends in the time series (linear, second-order, and third-order) were removed on a voxelwise basis to reduce low frequency noise components. Functional and anatomical images were transformed to common space (Tailairach and Tournoux, 1988), and the voxels resampled to 2 × 2 × 2 mm. Voxelwise normalization was performed by dividing the signal intensity at each time point by the voxel's mean intensity.

2.3.2.2.2. Statistical analysis. For each individual subject, a general linear modeling (GLM) approach was used to identify brain regions in which the time course of the BOLD signal was significantly related to the task, either the painful stimulus or the motor task (AFNI script 3dDeconvolve). The GLM for the pain scans consisted of three temporally independent regressors (one for each temperature of painful heat and for the VAS rating task) each represented by a delayed boxcar function convolved using the AFNI BLOCK function to account for hemodynamic delay. The GLM also included 6 motion correction parameters as regressors. Though the GLM included a regressor for the VAS rating task, results for this regressor are not presented because brain activity during the VAS rating task is not a variable of interest in this study. The GLM for the motor scans consisted of a regressor for the motor task as well as motion correction parameters. Voxelwise regression of the BOLD signal time course with the appropriate model resulted in statistical parametric maps for pain-related activation and motor-related activation for each individual subject.

Group maps of significant pain-related activation were calculated separately for each stimulus type (48 °C and 50VAS) using regression coefficients from the GLM, collapsed across sessions, in a one-sample *t*-test (AFNI routine 3dttest). The resulting group statistical parametric maps (one for the 48 °C stimulus and one for the 50VAS stimulus) were thresholded to identify significant activation associated with each type of painful heat using a cluster threshold approach to correct for multiple comparisons across the brain. The cluster threshold criterion was determined using Monte Carlo simulations to estimate the likelihood of detecting false positives over multiple comparisons (AFNI routine 3dClustSim). Based on the simulations, which were derived from whole-brain analyses, a significant cluster (overall corrected $p < 0.05$) was defined as a minimum cluster size of 190 mm^3 of contiguous voxels, each with a voxelwise $p < 0.05$. The cluster threshold criterion was applied to maps for the t-statistic for each stimulus type (48 °C and 50VAS) to identify the voxels that responded significantly to each type of painful heat, resulting in two thresholded maps for each subject: (1) voxels responding significantly to 48 °C stimuli and (2) voxels responding significantly to the 50VAS stimulus. Statistical parametric maps for the motor task were thresholded using the same approach to identify voxels responding significantly during the motor task. The pain-related and motor-related maps that resulted from this stage of analysis differed in spatial extent. Accordingly, cluster size thresholding for subsequent analyses (described below) involved different spatial threshold criteria for the pain- and motor-related activation maps.

To evaluate reliability of pain- and motor-related fMRI activation across the three sessions conducted in this study, two measures were calculated for voxels that showed significant pain-related activation, separately for each stimulus type (48 °C and 50VAS), or significant motor-related activation: (1) spatial reliability coefficients based on spatial localization and extent of activation and (2) intraclass correlation coefficients (ICC) based on the GLM regression coefficients, which reflect fMRI response amplitude. Spatial reliability coefficients are based on the number of voxels commonly activated in all sessions (Rombouts et al., 1998); voxels significantly activated by painful stimuli or the motor task were identified in each session using regression coefficients from the GLM in a one-sample *t*-test. Spatial reliability coefficients were calculated separately for the 48 °C stimulus, the 50VAS stimulus, and the motor task, using the formula R = (3 × number of voxels commonly activated in all 3 sessions)/(sum of activated voxels in all 3 sessions) (Havel et al., 2006). Spatial reliability coefficients for the painful stimuli were also calculated across the two stimulus types (48 °C and 50VAS) using the formula R = (2 × number of voxels commonly activated by both stimulus types)/(sum of activated voxels by both stimulus types). Spatial reliability coefficients can range in value from 0 to 1, with 0 indicating poor spatial reliability and 1 indicating perfect spatial overlap across sessions or conditions.

ICCs were calculated using the regression coefficients from the GLM (which reflect BOLD response amplitude) separately for the 48 °C stimulus, the 50VAS stimulus, and the motor task, resulting in three reliability maps (one for the 48 °C stimulus, one for the 50VAS stimulus, and one for the motor task). ICCs were calculated separately for each painful stimulus type and the motor task (AFNI routines 3dCalc and 3dMean) using the formula described in McGraw and Wong (1996) for the degree of absolute agreement among repeated measurements:

$$ICC(A,1)\frac{(BMS-EMS)}{BMS+(k-1)EMS+\frac{k}{n}(WMS-EMS)}$$

The ICC therefore measures the correlation of the magnitude of pain- or motor-related fMRI responses between sessions using a two-way mixed ANOVA framework, where the variance is divided into between subject variance (*BMS*), within subject variance (*WMS*), and residual sources of variance (*EMS*), *k* is the number of repeated sessions, and *n* is the number of subjects within a session. An F-statistic and *p*-value associated with each voxel's ICC were calculated based on the approach described in (McGraw and Wong, 1996). The F-statistic maps were then thresholded using a cluster threshold approach to correct for multiple comparisons across the voxels that survived the initial thresholding step (described above) (AFNI routine 3dClustSim). As previously noted, the pain- and motor-related maps differed in spatial extent after the initial thresholding step; as a result, different cluster size threshold criteria were calculated for these maps in these subsequent analyses. For pain-related maps, a significant cluster (overall corrected $p < 0.05$) was defined as a minimum cluster size of 127 mm^3 of contiguous voxels, each with a voxelwise $p < 0.05$. For motor-related maps, a significant cluster (overall corrected $p < 0.05$) was defined as a minimum cluster size of 148 mm^3 of contiguous voxels, each with a voxelwise $p < 0.05$. The thresholding step was conducted separately for each stimulus type and for the motor task, resulting in three statistical parametric maps (one

for the 48 °C stimulus, one for the 50VAS stimulus, and one for the motor task) of brain areas with significantly reliable pain- or motor-related activation. These maps were then masked to exclude voxels where within-subject variance contributed to more than 1% of the total variance (calculated using AFNI's 3dICC_REML routine). The purpose of this step was to eliminate voxels with artifactually high ICCs (such as would be the case if BMS was high, WMS was high, and EMS was low), thereby reducing the possibility of false positives (Chen et al., 2007). EMS represents random variance with unknown sources that might include MRI-related noise, physiological noise, or cognitive processes unrelated to the task that change over time (Bennett and Miller, 2010).

Significantly reliable pain-related activation (as determined by clusters with significant ICCs) associated with each stimulus type was examined in anatomically-defined regions of interest (ROI) known to be involved in processing painful stimuli: the arm representation area of S1, S2, pACC, aMCC (corresponding to the area referred to as the mid-ACC in many previous pain studies), anterior INS (aINS), posterior INS (pINS), SMA, inferior frontal gyrus (IFG), medial prefrontal cortex (mPFC) and dorsolateral prefrontal cortex (dlPFC). The arm area of S1 was defined as the region of the postcentral gyrus starting from the most medial portion of the hand representation (delineated by the "knob" created by the postcentral gyrus) and extending approximately 2 cm medially along the surface from that point, excluding digit representations (Servos et al., 1998; van Westen et al., 2004); these boundaries should encompass the complete arm representation. The boundaries of the pACC and aMCC were delineated based on Vogt (2005). The boundaries of the other ROIs were described previously (Moulton et al., 2005). Significantly reliable motor-related activation (as determined by clusters with significant ICCs) was examined in two anatomically-defined motor ROIs: the hand representation area in the primary motor cortex (M1) and SMA.

For the painful stimuli, separate evaluations were conducted for portions of each ROI contralateral (right hemisphere) and ipsilateral (left hemisphere) to the stimuli. For each stimulus type, the largest cluster of significantly reliable (as defined by significant ICC values) pain-related activation was identified in each ROI. The reliability of each cluster was then classified using the peak ICC value based on the conservative criteria described by Shrout (1998): virtually no reliability (0–0.1), slight reliability (0.11–0.4), fair reliability (0.41–0.6), moderate reliability (0.61–0.8), and substantial reliability (0.81–1). The motor data were analyzed using the same approach, with the largest cluster of significantly reliable motor-related activation identified in each ROI contralateral to the hand performing the task (left hemisphere) and ipsilateral (right hemisphere).

To address the question of whether a painful stimulus of constant temperature or of constant perceived intensity produced more reliable BOLD fMRI responses, a group-level contrast between significantly reliable pain-related activation for the 2 stimulus types was conducted in which the voxelwise ICC values for each stimulus type were contrasted (Donner and Zou, 2002).

2.3.2.2.3. Additional statistical analyses

2.3.2.2.3.1. Perception-Dependent Responses. To address the question of whether the perceived pain intensity-dependent response is consistent across sessions, a voxelwise analysis was conducted in which individual pain intensity ratings for each subject and each stimulus were used in the GLM. The analysis was conducted on the 48 °C pain condition, which was the protocol in which ratings varied the most across subjects. For each individual subject and each session, pain intensity ratings of each 48 °C stimulus were used in the regression model for each voxel, resulting in parameter estimates for each subject at each voxel that represented the degree to which perceived pain intensity covaried with the BOLD response. The parameter estimates were then used in a group analysis to identify voxels where the pain intensity ratings significantly predicted the magnitude of the BOLD response to the stimulus. Intersession reliability in these voxels could then be evaluated by calculating ICCs.

2.3.2.2.3.2. Intersession Reliability: The effect of duration between sessions. To address the question of whether the duration between sessions had an effect on intersession reliability, we reanalyzed our pain and motor task data sets to calculate ICCs on task-related activation detected when duration between session was used as a covariate in the analysis. The analysis involved two stages.

The first analysis stage involved conducting a voxelwise GLM repeated measures ANCOVA for each of our pain and motor tasks separately, using each subject's average days between sessions as a time-invariant covariate using the methodology of Winer (1971). The analysis excluded voxels that violated the assumption of homogeneity of regression slopes because ANCOVA is not an appropriate statistical test when this assumption is violated (Tabachnick and Fidell, 2007). Our homogeneity test revealed that 98% of voxels across the brain did NOT violate this assumption and were therefore included in the ANCOVA.

In the second stage of the analysis, ICCs were calculated based on the ANOVA framework described by McGraw and Wong (1996), but with a slight modification: A two-way mixed design of absolute agreement was used to calculate ICCs on a voxelwise basis using adjusted variability estimates and degrees of freedom derived from the ANCOVA analysis (AFNI routines 3dcalc and 3dMean). Subsequent analysis steps involved the same thresholding, masking, and statistical steps as the original ICC analysis.

2.3.3. Statistical Analysis: Psychophysical data

Pain intensity ratings obtained in the scanner were evaluated separately for each stimulus type using the nonparametric Friedman test for intersession effects. Mauchly's test of sphericity was used to compare the error variance associated with each stimulus type. The significance level for all tests was set at 0.05.

3. Results

3.1. Psychophysics

Pain intensity ratings did not differ significantly across sessions for either stimulus type ($p > 0.05$, Friedman test, Fig. 1), indicating that within the parameters of this study, repeated testing did not change perceived intensity of the painful heat stimuli. Variance of pain intensity ratings was greater for the 48 °C stimulus than the perceptually-equalized 50VAS stimulus ($p < 0.05$).

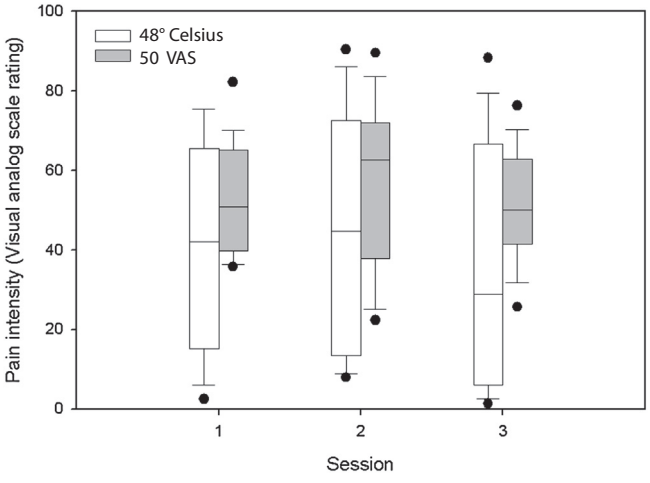

Fig. 1. Box plots of pain intensity ratings for 48 °C stimulus and perceptually-equalized stimuli (50VAS) where the temperature that produced a rating of 50 on a 0–100 visual analog scale was selected individually for each subject ($n = 14$). Ratings were obtained in the scanner on separate days (sessions 1–3). Median values are represented by solid lines. Solid circles represent individual outliers. No significant session effects were found for either measure (Friedman test, $p > 0.05$ for each stimulus type).

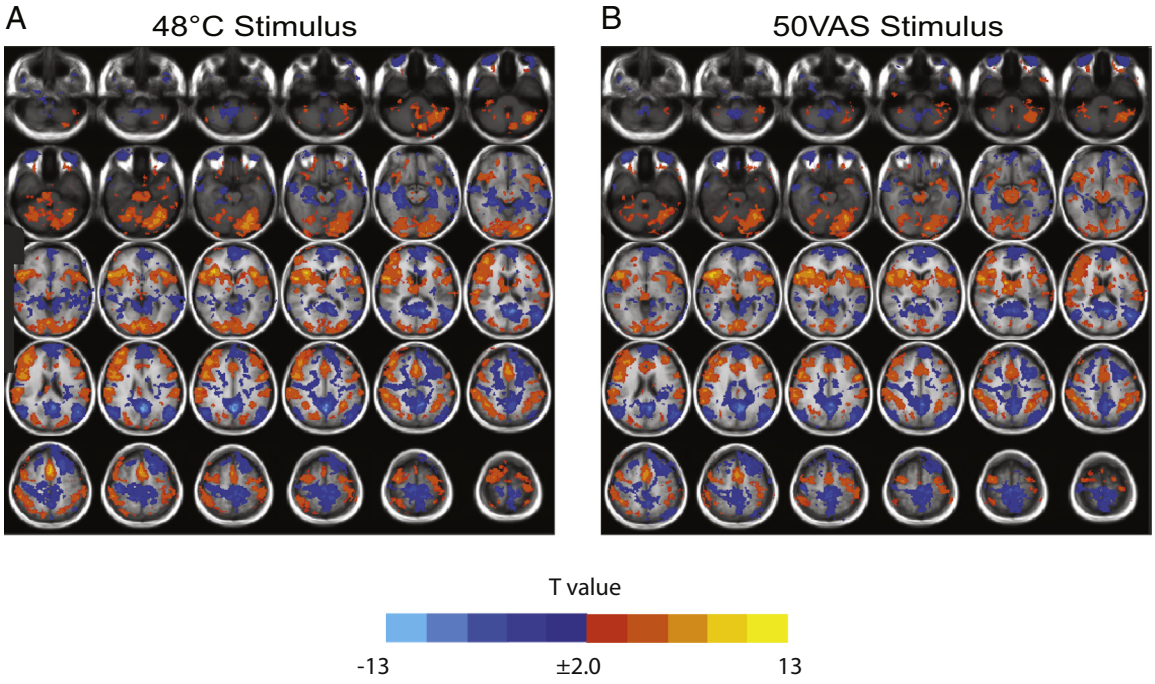

Fig. 2. Group maps of brain regions significantly activated by painful heat stimuli: (A) 48 °C and (B) subject-specific temperatures that were rated as 50 on a 0–100 visual analog scale for pain intensity (50VAS). Functional activation is overlaid on the T1-weighted group average of each subject's brain normalized to Talairach space. Significant activation was defined as a minimum cluster size of 9 contiguous voxels (190 mm^3), each with a voxelwise $p < 0.05$, resulting in an overall corrected $p < 0.05$. Orange and yellow areas represent voxels with a significantly positive pain-related BOLD response, while blue areas represent voxels with a significantly negative pain-related BOLD response.

3.2. Pain-related Activation

Significant pain-related activation was found in expected brain regions of interest (ROI) for both types of painful heat stimuli (Fig. 2). Most prescribed ROIs showed significant increases in the BOLD signal associated with painful heat, including pACC, aMCC, aINS, and S2. A few ROIs, including the arm region of S1 and frontal lobe regions, showed significant decreases in the BOLD signal in response to painful stimuli (Fig. 2). The map of pain-related activation from this study was compared with the reverse inference meta-analysis image from the Neurosynth database (Yarkoni et al., 2011) and was found to have notable overlap, particularly with respect to the insula, anterior cingulate cortex, and some dorsolateral frontal regions. As reported above, robust activation was found in several other brain regions that were not identified in the image from the Neurosynth database, but have been described in the literature (Duerden and Albanese, 2013).

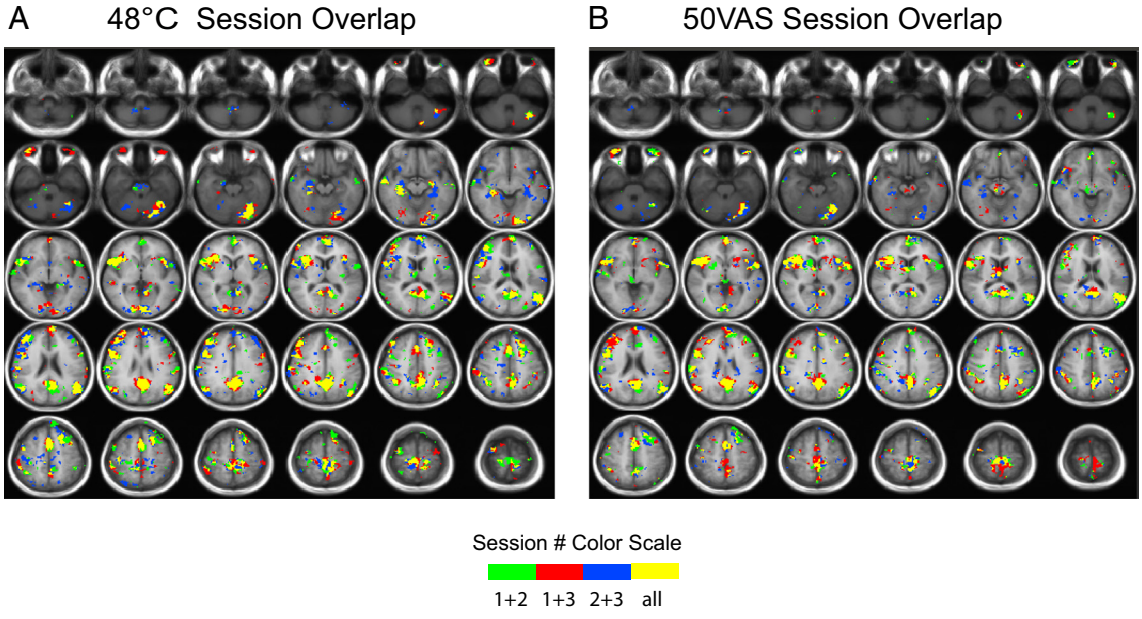

Fig. 3. Group maps showing voxels significantly activated in 1, 2, or all 3 sessions in response to painful heat stimuli: (A) 48 °C and (B) subject-specific temperatures that were rated as 50 on a 0–100 visual analog scale for pain intensity (50VAS).

Table 1
Spatial reliability coefficients for pain-activated regions associated with 48 °C stimulus.

Region of interest[a]	Number of significant voxels (each session)			Across-session overlap (# voxels)	Reliability coefficient[b]
	1	2	3		
Anterior midcingulate cortex (left)	1352	1592	1976	472	0.29
Anterior midcingulate cortex (right)	1888	1856	2768	496	0.23
Pregenual anterior cingulate cortex (left)	128	168	432	0	0.00
Pregenual anterior cingulate cortex (right)	120	264	128	0	0.00
Inferior frontal gyrus (left)	3504	2600	3672	496	0.15
Inferior frontal gyrus (right)	5472	8256	8848	3408	0.45
Primary somatosensory cortex (left)	592	104	624	8	0.02
Primary somatosensory cortex (right)	160	664	408	16	0.04
Supplementary motor area (left)	1360	1632	2264	160	0.09
Supplementary motor area (right)	2056	1936	2088	400	0.20
Anterior insular cortex (left)	1992	3648	2544	1112	0.41
Posterior insular cortex (left)	360	768	224	0	0.00
Anterior insular cortex (right)	3280	3736	2936	1744	0.53
Posterior insular cortex (right)	384	136	312	0	0.00
Medial prefrontal cortex (left)	3816	5320	5688	1904	0.39
Medial prefrontal cortex (right)	2640	2944	2400	624	0.23
Dorsolateral prefrontal cortex (left)	6928	8136	5496	976	0.14
Dorsolateral prefrontal cortex (right)	10,352	12,464	13,824	4672	0.38
Second somatosensory cortex (left)	1088	1264	1376	112	0.09
Second somatosensory cortex (right)	880	1872	1368	384	0.28
Thalamus (left)	560	360	576	24	0.05
Thalamus (right)	256	1656	392	72	0.09

[a] Left regions are ipsilateral to the stimulus
[b] Reliability coefficient = (3 × number of common voxels)/(sum of activated voxels in each session)

3.2.1. Reliability of pain-related activation: Spatial reliability coefficients

Intersession reliability of the spatial extent of significant pain-related activation is shown in Fig. 3. Spatial reliability coefficients were low (< 0.2) in most pain-related ROIs, for both the 48 °C stimulus (Table 1) and 50VAS stimulus (Table 2). Spatial reliability differed by ROI, with the highest spatial overlap shown for both stimulus types in the aINS and no voxels displaying complete overlap (e.g., significant activation in all 3 sessions) for either stimulus type in the pACC.

Spatial overlap between the two stimulus types (48 °C and 50VAS) is shown in Fig. 4. Spatial reliability coefficients were relatively high (most > 0.5, Table 3), indicating moderately consistent spatial extent of activation regardless of whether the stimulus was of constant temperature or of constant perceived pain intensity. Spatial overlap between the stimulus types showed regional differences, with the highest reliability coefficients in the aINS and low reliability coefficients in the pACC.

3.2.2. Reliability of pain-related activation: Intraclass correlation coefficients

Reliability of pain-related activation amplitude was assessed in a two-step process (see Methods), by calculating voxelwise ICCs and then applying a statistical filter to eliminate voxels with artifactually high ICC values due to a combination of high WMS and low EMS. Fig. 5 shows an example of the importance of the statistical filtering step. Fig. 5A shows the average BOLD response amplitude for each

Table 2
Spatial reliability coefficients for pain-activated regions associated with 50VAS[a]**stimulus**.

Region of interest[b]	Number of significant voxels (each session)			Across-session overlap (# voxels)	Reliability coefficient[c]
	1	2	3		
Anterior midcingulate cortex (left)	1504	1408	928	168	0.13
Anterior midcingulate cortex (right)	1392	2152	1904	424	0.23
Pregenual anterior cingulate cortex (left)	344	392	0	0	0.00
Pregenual anterior cingulate cortex (right)	176	96	64	0	0.00
Inferior frontal gyrus (left)	3800	3928	3880	440	0.11
Inferior frontal gyrus (right)	6760	7192	6512	3288	0.48
Primary somatosensory cortex (left)	248	240	184	8	0.04
Primary somatosensory cortex (right)	32	72	400	0	0.00
Supplementary motor area (left)	1688	1088	1528	168	0.12
Supplementary motor area (right)	1744	1512	1576	344	0.21
Anterior insular cortex (left)	2384	3032	3592	984	0.33
Posterior insular cortex (left)	136	176	888	8	0.02
Anterior insular cortex (right)	4344	2784	4608	2200	0.56
Posterior insular cortex (right)	104	296	224	0	0.00
Medial prefrontal cortex (left)	4648	3888	2608	776	0.21
Medial prefrontal cortex (right)	2416	1808	2104	480	0.23
Dorsolateral prefrontal cortex (left)	5960	6544	3944	632	0.12
Dorsolateral prefrontal cortex (right)	7144	4088	8800	1136	0.17
Second somatosensory cortex (left)	336	1160	736	24	0.03
Second somatosensory cortex (right)	728	1800	552	208	0.20
Thalamus (left)	168	904	352	0	0.00
Thalamus (right)	1664	1656	1368	200	0.13

[a] 50VAS is the subject-specific temperature that produced a perceived intensity of 50 on a 0–100 visual analog scale
[b] Left regions are ipsilateral to the stimulus
[c] Reliability coefficient = (3 × number of common voxels)/(sum of activated voxels in each session)

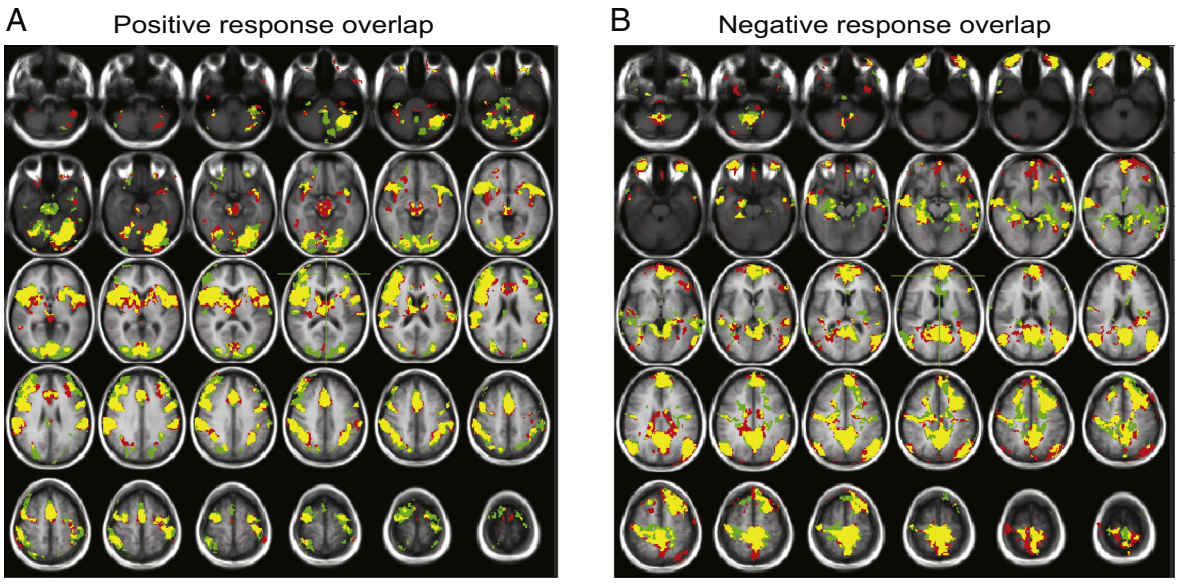

Fig. 4. Overlap of significant pain related activation (A) or deactivation (B) produced by 48 °C stimuli (green areas) and subject-specific temperatures that were rated as 50 on a 0–100 visual analog scale for pain intensity (50VAS, red areas). Yellow represents areas that showed significant response for both stimulus conditions.

session in the voxel in the left IFG with the highest ICC prior to the statistical filtering step. The graph shows a highly variable pain-related BOLD response in this voxel across sessions; however, the ICC value (0.75) calculated for this voxel was the highest in the ROI and statistically significant. The high ICC value obtained for this voxel is attributable to a combination of high BMS, high WMS and low EMS. This example illustrates that it is possible to obtain a high ICC value for a voxel despite high variability (low reliability) of the response from session to session. Thus, relying entirely on ICC values without examining the variances that contribute to the ICC can result in a voxel being identified as having a reliable intersession response when in fact it does not (in other words, a false positive). The statistical filtering step employed here, which eliminates voxels with high WMS (such as the voxel shown in Fig. 5A), is a conservative approach to address this issue. Fig. 5B shows the average BOLD response for each session in the voxel in the left IFG with the highest ICC *after* the statistical filtering step. This voxel, which is anatomically close to the peak voxel found prior to statistical filtering, shows much more consistent and reliable pain-related responses across session, and an ICC value (0.73) that more accurately reflects the low intersession variance.

Intersession reliability of pain-related activation associated with the 48 °C stimulus is summarized in Table 4 and Fig. 6A. Table 4 identifies the largest cluster of significantly reliable pain-related activation found in each ROI and the ICC value associated with the peak voxel in the cluster; additional clusters were also found in most ROIs. Fair-to-moderate reliability (based on Shrout, 1998) was found in every

Table 3
Spatial reliability coefficients for pain-activated regions across stimulus conditions.

Region of interest[a]	48 °C stimulus (# voxels)	50 VAS[b] stimulus (# voxels)	Overlap (# voxels)	Reliability coefficient[c]
Anterior midcingulate cortex (left)	3520	3960	2512	0.672
Anterior midcingulate cortex (right)	4016	4296	3080	0.741
Pregenual anterior cingulate cortex (left)	512	560	192	0.358
Pregenual anterior cingulate cortex (right)	600	392	192	0.387
Inferior frontal gyrus (left)	8400	9304	5800	0.655
Inferior frontal gyrus (right)	13,448	11,120	10,240	0.834
Primary somatosensory cortex (left)	920	592	448	0.593
Primary somatosensory cortex (right)	1144	400	312	0.404
Supplementary motor area (left)	3976	3784	2632	0.678
Supplementary motor area (right)	4264	3416	2680	0.698
Anterior insular cortex (left)	5104	5776	4856	0.893
Posterior insular cortex (left)	1224	848	360	0.347
Anterior insular cortex (right)	5600	6544	5336	0.879
Posterior insular cortex (right)	456	408	72	0.167
Medial prefrontal cortex (left)	9544	7560	6968	0.815
Medial prefrontal cortex (right)	5672	5496	4000	0.716
Dorsolateral prefrontal cortex (left)	15,560	13,552	9824	0.675
Dorsolateral prefrontal cortex (right)	23,024	16,088	14,712	0.752
Second somatosensory cortex (left)	2936	2576	2008	0.729
Second somatosensory cortex (right)	3056	2584	2352	0.834
Thalamus (left)	1192	952	392	0.366
Thalamus (right)	2040	2776	1584	0.658

[a] Left regions are ipsilateral to the stimulus
[b] 50VAS is the subject-specific temperature that produced a perceived intensity of 50 on a 0–100 visual analog scale
[c] Reliability coefficient = (3 × number of common voxels)/(sum of activated voxels in each session)

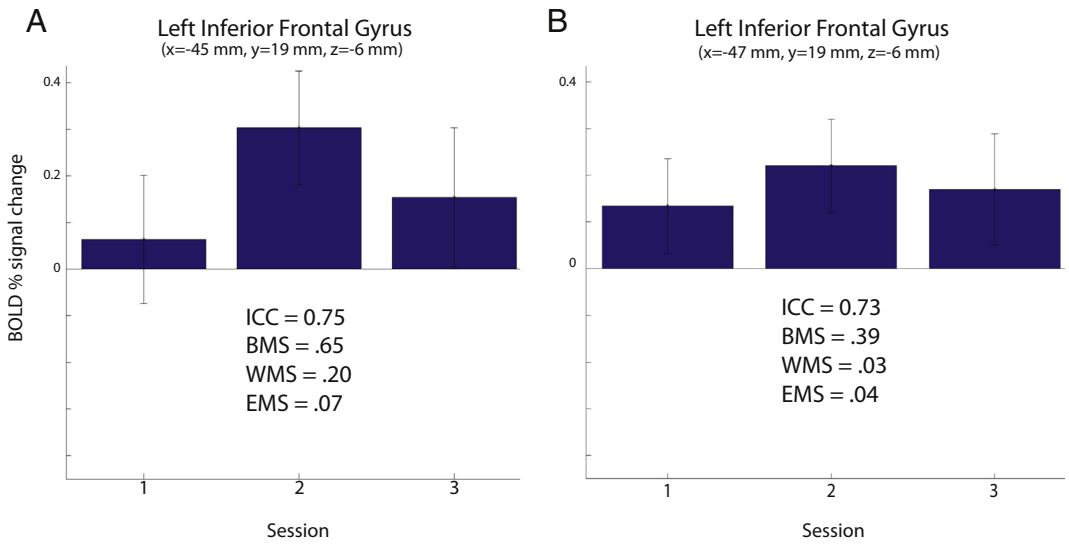

Fig. 5. BOLD fMRI response amplitude in the left inferior frontal gyrus to painful heat stimuli rated 50 on a 0–100 visual analog scale for pain intensity (50VAS). (A) Responses in the voxel with the peak intraclass correlation coefficient (ICC) value before statistical filtering. (B) Responses in the voxel with the peak ICC after filtering. Values for ICC, between-subject variance (BMS), within-subject variance (WMS), and residual error variance (EMS) are shown for each voxel. Statistical filtering consisted of removing voxels where WMS contributed to more than 1% of the total variance. The value of statistical filtering is demonstrated in (A) where the voxel with the peak ICC value has high WMS, indicating low reliability, but an artifactually high ICC due to high BMS and low EMS. Statistical filtering eliminated this voxel, instead identifying a nearby voxel (B) with a peak ICC value and low WMS in the same brain region of interest.

ROI except (1) contralateral pACC, which showed slight reliability, and (2) S1, pINS, and ipsilateral thalamus, which contained no clusters that were significantly reliable across the three sessions. ROIs with the highest ICCs (the upper portion of the moderate range) included aMCC, aINS, and several frontal lobe areas. Clusters of significantly reliable pain-related activation from selected ROIs are shown in Fig. 6A.

Intersession reliability of pain-related activation associated with the 50VAS stimulus is summarized in Table 5 and Fig. 6B. Table 5 identifies the largest cluster of significantly reliable pain-related activation found in each ROI and the ICC value associated with the peak voxel in the cluster; additional clusters were also found in most ROIs. Fair-to-moderate reliability (based on Shrout, 1998) was found in every ROI except (1) ipsilateral pACC and ipsilateral thalamus, which showed slight reliability, and (2) contralateral pACC, S1, and contralateral pINS, which contained no clusters that were significantly reliable across the three sessions. ROIs with the highest ICCs (the upper portion of the moderate range) included aMCC and S2. Clusters of significantly reliable pain-related activation from selected ROIs are shown in Fig. 6B.

The results of the group-level contrast between significantly reliable pain-related activation for the 2 stimulus types are summarized in Table 6. Voxelwise ICC values for each stimulus type were contrasted to obtain these results. The constant temperature stimulus (48 °C) produced more reliable activation than the constant perceived pain intensity stimuli in aINS (bilaterally), ipsilateral mPFC, and ipsilateral dlPFC; the converse was found in contralateral IFG, contralateral dlPFC, and ipsilateral S2. No significant differences in reliability were found for the other ROIs examined in this study. Thus, though there were some differences, no consistent pattern emerged in terms of whether a painful stimulus of constant temperature or of constant perceived intensity produced more reliable BOLD fMRI responses.

3.2.3. Perception-dependent responses

Clusters for which pain intensity ratings of the 48 °C stimulus significantly predicted the magnitude of the BOLD response to the stimulus were found to be very limited, scattered, and with little overlap across

Table 4
Brain regions with significantly reliable pain-related signal amplitude associated with 48 °C stimulus.

Region of interest[a]	Largest cluster (mm^3)	Peak voxel in largest cluster					
		x	*y*	*z*	ICC	*F*	*p*-value
Anterior midcingulate cortex (left)	1144	−1	23	32	0.76	9.908	< 0.001
Anterior midcingulate cortex (right)	704	1	23	28	0.746	9.314	< 0.001
Pregenual anterior cingulate cortex (left)	56	−3	43	10	0.497	3.992	0.001
Pregenual anterior cingulate cortex (right)	80	1	37	10	0.39	2.808	0.012
Inferior frontal gyrus (left)	392	−47	17	−4	0.634	6.081	< 0.001
Inferior frontal gyrus (right)	3096	57	9	22	0.757	10.359	< 0.001
Supplementary motor area (left)	288	−1	−13	54	0.683	7.048	< 0.001
Supplementary motor area (right)	704	13	−11	62	0.593	5.372	< 0.001
Anterior insular cortex (left)	528	−33	15	6	0.668	6.95	< 0.001
Anterior insular cortex (right)	2232	37	5	−6	0.721	8.749	< 0.001
Medial prefrontal cortex (left)	1896	−5	49	10	0.688	7.653	< 0.001
Medial prefrontal cortex (right)	1376	9	59	8	0.771	10.632	< 0.001
Dorsolateral prefrontal cortex (left)	816	−33	21	40	0.63	5.899	< 0.001
Dorsolateral prefrontal cortex (right)	640	45	1	38	0.734	8.967	< 0.001
Second somatosensory cortex (left)	184	−55	−25	16	0.451	3.506	0.003
Second somatosensory cortex (right)	1152	55	−25	18	0.657	6.722	< 0.001
Thalamus (right)	32	7	−11	14	0.436	3.169	0.006

[a] Left regions are ipsilateral to the stimulus

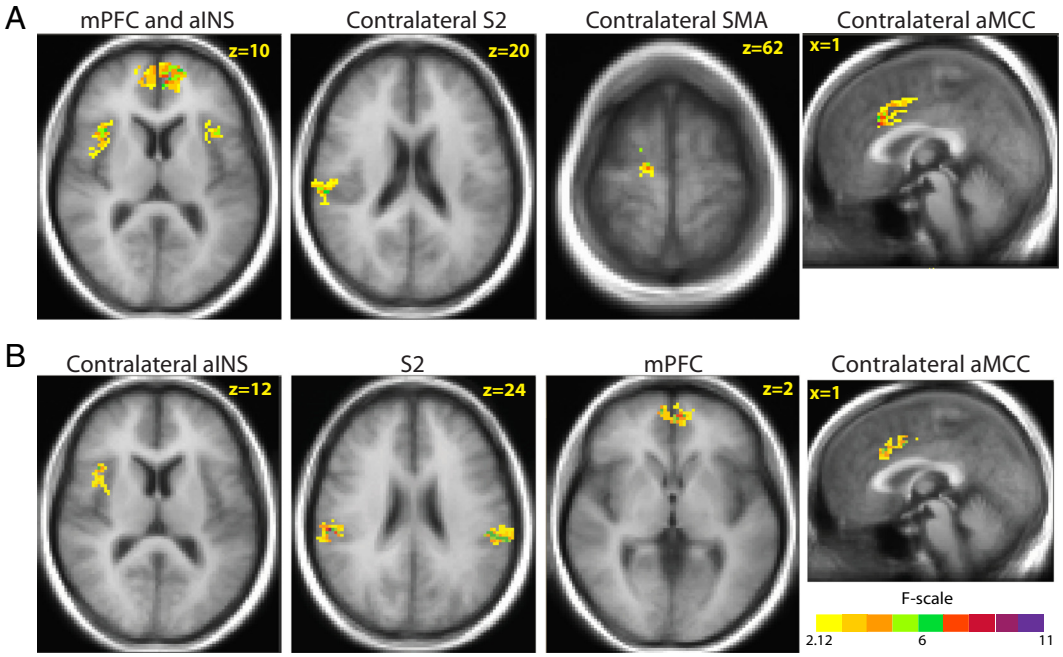

Fig. 6. Representative brain regions with significantly reliable responses to (A) painful 48 °C heat stimuli and (B) painful heat stimuli rated 50 on a 0–100 visual analog scale for pain intensity across 3 sessions conducted on separate days. Color-coded areas represent the largest clusters in a brain region of interest with statistically significant intraclass correlation coefficients (threshold criterion of 6 contiguous voxels equivalent to 127 mm^3, each with a voxelwise $p < 0.05$, resulting in overall corrected $p < 0.05$). Regions of interest with significantly reliable pain-related activation included the medial prefrontal cortex (mPFC), anterior insular cortex (aINS), second somatosensory cortex (S2), anterior mid-cingulate cortex (aMCC), and supplementary motor area (SMA).

sessions (Supplemental Fig. 1). As a result, a quantitative analysis of ICCs was deemed unnecessary.

3.2.4. Effect of duration between sessions on intersession reliability

As shown in Supplemental Tables 1, 2, and 3, the effect of duration between sessions on the reliability of pain- and motor-related fMRI activation was minor. The ICCs calculated in our original analysis and the analysis that took between-session duration into account are comparable and did not differ in a large or consistent way in any pain- or motor-related region of interest. Thus, the results suggest that with this data set, the duration between sessions did not contribute significantly to intersession variability.

3.3. Motor activation

Significant motor-related activation was found in expected brain ROIs (M1 and SMA) bilaterally. Intersession reliability of the spatial extent of significant motor-related activation is shown in Table 7. The spatial reliability coefficient was high in M1 contralateral to the hand performing the motor task (0.78) but was much lower in contralateral SMA (0.2). The spatial reliability coefficients in M1 and SMA ipsilateral to the hand performing the motor task were notably lower than the contralateral regions. Intersession reliability of the amplitude of motor-related activation is shown in Table 7, which identifies the largest cluster of significantly reliable motor-related activation in each ROI and the ICC value associated with the peak voxel in the cluster;

Table 5
Brain regions with significantly reliable pain-related signal amplitude associated with 50VAS[a] stimulus.

Region of interest[b]	Largest cluster (mm^3)	Peak voxel in largest cluster					
		x	*y*	*z*	ICC	*F*	*p*-value
Anterior midcingulate cortex (left)	1296	−3	21	32	0.694	7.823	< 0.001
Anterior midcingulate cortex (right)	888	3	−1	40	0.781	10.967	< 0.001
Pregenual anterior cingulate cortex (left)	48	−11	33	22	0.392	2.898	0.010
Inferior frontal gyrus (left)	384	−49	11	28	0.596	5.497	< 0.001
Inferior frontal gyrus (right)	2816	45	21	10	0.708	8.007	< 0.001
Supplementary motor area (left)	280	−1	−11	58	0.675	6.874	< 0.001
Supplementary motor area (right)	280	1	−9	58	0.613	5.507	< 0.001
Anterior insular cortex (left)	336	−47	5	−2	0.531	4.271	< 0.001
Posterior insular cortex (left)	40	−37	−29	20	0.457	3.449	0.003
Anterior insular cortex (right)	832	39	17	12	0.636	6.035	< 0.001
Medial prefrontal cortex (left)	584	−3	53	2	0.706	7.879	< 0.001
Medial prefrontal cortex (right)	328	5	55	0	0.69	7.33	< 0.001
Dorsolateral prefrontal cortex (left)	304	−25	11	56	0.619	5.593	< 0.001
Dorsolateral prefrontal cortex (right)	2200	43	41	22	0.696	7.537	< 0.001
Second somatosensory cortex (left)	1688	−57	−31	20	0.774	10.818	< 0.001
Second somatosensory cortex (right)	768	55	−25	24	0.731	9.296	< 0.001
Thalamus (left)	16	−5	−3	8	0.289	2.187	0.043
Thalamus (right)	216	7	−3	8	0.564	4.649	< 0.001

[a] 50VAS is the subject-specific temperature that produced a perceived intensity of 50 on a 0–100 visual analog scale
[b] Left regions are ipsilateral to the stimulus

Table 6
Brain regions with significant differences in reliability[a] of pain-related signal amplitude across stimulus conditions.

Region of interest[b]	Stimulus type[c]	Largest cluster (mm^3)	Z stat	x	Y	z	ICC	F	*p*-value	Reliability result
Inferior frontal gyrus (right)	48 °C	128	−3.87	43	19	−8	−0.308	0.327	0.98065	50VAS > 48 °C
	50VAS	128	−3.87	43	19	−8	0.312	2.313	0.03334	
Anterior insular cortex (left)	48 °C	296	3.31	−37	15	10	0.61	5.414	0.00013	48 °C > 50VAS
	50VAS	296	3.31	−37	15	10	−0.112	0.681	0.76309	
Anterior insular cortex (right)	48 °C	208	4.21	35	9	−4	0.451	3.782	0.00189	48 °C > 50VAS
	50VAS	208	4.21	35	9	−4	−0.346	0.283	0.98975	
Medial prefrontal cortex (left)	48 °C	1000	5.00	−13	57	8	0.689	7.653	0.00001	48 °C > 50VAS
	50VAS	1000	5.00	−13	57	8	−0.259	0.406	0.95425	
Dorsolateral prefrontal cortex (left)	48 °C	288	3.64	−29	23	44	0.306	2.295	0.03459	48 °C > 50VAS
	50VAS	288	3.64	−29	23	44	−0.315	0.313	0.98401	
Dorsolateral prefrontal cortex (right)	48 °C	152	−2.83	27	23	46	0.044	1.13	0.37983	50VAS > 48 °C
	50VAS	152	−2.83	27	23	46	0.637	6.462	0.00003	
Second somatosensory cortex (left)	48 °C	528	−4.03	−49	−25	22	−0.146	0.622	0.81371	50VAS > 48 °C
	50VAS	528	−4.03	−49	−25	22	0.684	7.065	0.00001	

[a] Based on voxelwise contrast of significant intraclass correlation coefficients (48 °C stimulus versus 50VAS stimulus)
[b] Left regions are ipsilateral to the stimulus
[c] 50VAS is the subject-specific temperature that produced a perceived intensity of 50 on a 0–100 visual analog scale

additional clusters were also found in these ROIs. Substantial reliability (based on Shrout, 1998) was found in contralateral M1 (ICC = 0.815) and moderate reliability in contralateral SMA (ICC = 0.657). Fair reliability was found in M1 and SMA ipsilateral to the hand performing the motor task. Overall, motor-related activation showed greater reliability than pain-related activation for both the spatial extent and ICC measures.

4. Discussion

This study quantitatively evaluated across-session reliability of pain-related BOLD fMRI responses in brain areas that are typically regarded as part of the cortical pain network. While these brain areas show consistent and often robust activation in fMRI studies across diverse pain modalities (as reviewed in Duerden and Albanese, 2013), the question of how reliable this activation is from session to session has not been comprehensively addressed in the literature. We used two reliability measures (voxelwise spatial overlap and ICCs based on BOLD response amplitude) that are commonly used in the literature (Bennett and Miller, 2010) to assess test–retest reliability of fMRI responses to a wide variety of conditions (e.g., motor tasks, visual stimulation, memory) but have not been previously applied to pain. The results revealed that the two measures of intersession pain-related fMRI activation reliability used in this study produced disparate results, with (1) reliability based on spatial measures generally low and highly variable across brain regions and (2) reliability based on signal amplitude (ICCs) in the fair-to-moderate range for most brain regions.

4.1. Spatial measures of intersession reliability

Spatial reliability coefficients were low for most regions examined, indicating a low probability that the same voxels are activated across three separate sessions. The highest spatial reliability coefficients were found in the aINS for both stimulus types but no spatial overlap across the three sessions was found for pINS. For both stimulus types, spatial reliability coefficients were higher in S2 than in S1 or thalamus, which is consistent with the findings of Taylor and Davis (2009), who used a different measure of spatial reliability in a study involving mechanical pain applied across 4 sessions in 6 subjects; their analysis was limited to S1, S2, and thalamus. While low S1 spatial reliability is in agreement with the inconsistency of S1 activation across fMRI studies of pain (Bushnell et al., 1999), the reasons for a total lack of spatial overlap in pACC are unclear. These results suggest that the precise location of pain-related activation is not highly reproducible across multiple sessions, perhaps due in part to variability introduced by motion correction, spatial normalization, and spatial smoothing procedures. As noted by Taylor and Davis (2009), conservative thresholding procedures such as those used in this study to avoid false positives may add to spatial variability across sessions by increasing false negatives; for example, a voxel may be activated by pain in all sessions but be eliminated from

Table 7
Reliability measures for motor task activation.

Spatial reliability coefficients for motor task-activated regions

Region of interest[b]	Number of significant voxels (each session)			Across-session overlap (# voxels)	Reliability coefficient[a]
	1	2	3		
Primary motor cortex (left)	2656	3608	3640	2576	0.78
Primary motor cortex (right)	80	256	792	16	0.04
Supplementary motor area (left)	1840	1160	2160	352	0.20
Supplementary motor area (right)	1368	1816	1784	16	0.01

Brain regions with significantly reliable motor task-related signal amplitude

Region of interest[b]	Largest cluster (mm^3)	Peak voxel in largest cluster					
		x	*y*	*z*	ICC	*F*	*p*-value
Primary motor cortex (left)	1984	−33	−21	48	0.815	15.58	< 0.001
Primary motor cortex (right)	152	29	−25	50	0.443	3.366	.004
Supplementary motor area (left)	760	−1	−9	54	0.657	7.116	< 0.001
Supplementary motor area (right)	104	17	−3	62	0.559	5.794	< 0.001

Brain regions with significantly reliable motor task-related signal amplitude
[a] Reliability coefficient = (3 × number of common voxels)/(sum of activated voxels in each session)
[b] Left regions are contralateral to the hand performing the task

one or more sessions if the activation falls slightly below threshold. As a result, spatial measures do not appear to be ideal for assessing intersession reliability because the threshold criterion has a major influence upon cluster sizes. However, suprathreshold clusters of voxels in a brain region may show reliable pain-related BOLD signal amplitude changes across sessions. To test this possibility, ICCs were calculated based on pain-related BOLD signal amplitude changes.

4.2. BOLD signal amplitude measures of intersession reliability

ICC calculation is a widely-accepted and commonly used method to evaluate test–retest reliability (Cacares et al., 2009; McGraw and Wong, 1996; Shrout and Fleiss, 1979). This study used a conservative two-step approach that involved calculating ICCs based on BOLD signal amplitude changes in suprathreshold voxels and then applying a filtering step to eliminate voxels with artifactually high ICCs based on the variance structure of the data. This approach revealed fair-to-moderate intersession reliability of pain-related activation in most regions of the cortical pain network, based on the conservative classification criteria described by Shrout (1998). Regional differences in intersession reliability were found, with the aMCC (bilaterally) containing clusters with the highest ICCs (≥ 0.7, with moderate reliability defined as 0.61–0.8) for both stimulus types. Other areas with ICCs in the moderate range included the aINS and most frontal lobe areas for both stimulus types, as well as S2 for the 50VAS stimulus. Areas with the lowest intersession reliability based on the ICC analysis also showed no to very low spatial reliability; these regions included pACC, S1, and pINS for both stimulus types. The finding of higher ICCs in S2 than S1 is consistent with Taylor and Davis (2009), though their calculated ICCs were higher than those in this study, likely because they included responses from subthreshold voxels in their analysis.

4.3. Regional differences in pain-related intersession reliability

The brain regions that displayed the highest intersession reliability in this study were aINS (based on both spatial extent and ICC measures) and aMCC (based on ICCs). Both of these areas have been implicated in processing affective aspects of pain, and are anatomically and functionally connected (Berthier et al., 1988; Friedman et al., 1986; Heimer and Van Hoesen, 2006; Vogt, 2005). Craig (2002, 2003, 2009, 2011) has postulated that the aINS has a role in generating subjective emotional feelings about the internal state of the body and interacts with the ACC, which initiates adaptive behavioral responses. Evidence for the involvement of aMCC in emotion-based response selection, including fear avoidance behavior (Vogt, 2005), provides support for Craig's model. Thus, the consistently reliable activation of aINS and aMCC in this study may highlight the importance of and priority given to processing emotional-motivational aspects of pain. However, Craig's model identifies pINS as the region that generates the initial cortical representation of the body's homeostatic condition, providing aINS with information upon which to generate emotions associated with that representation. The fact that the pINS had very low intersession reliability in this study does not fit well with a model of a serial-processing insular system in which the anterior region response is highly dependent upon the posterior region response.

The relatively high ICCs and spatial overlap for contralateral S2 contrasts sharply with that of the pINS. Both regions have historically been considered important in nociceptive processing, and most functional neuroimaging studies do not describe differences in their nociceptive processing capacities. The current study reveals the much more reliable responsiveness of the S2 cortex to acute heat stimuli, compared to the pINS. This difference suggests that S2 has a more essential role in nociceptive processing, at least within the context of responding to repetitively administered acute noxious heat stimuli.

4.4. Comparison of intersession reliability between stimulus types

The second objective of this study was to evaluate whether BOLD fMRI activation produced by painful heat stimuli of constant temperature or of constant perceived pain intensity was more reliable. As expected, spatial overlap of significant pain-related activation between these two conditions was relatively high, indicating that the experience of heat pain, regardless of these differences in intensities, activates common brain areas. This is consistent with a recent report that pain-related fMRI activation did not differ for fixed temperature stimuli and perceptually equalized stimuli in any brain region, including the pACC, aMCC, and INS (van den Bosch et al., 2013). The ICC contrast between the two stimulus conditions showed only a few brain regions where reliability differed (aINS, S2, and some frontal lobe regions). Among these brain regions, no clear pattern emerged: in some areas the constant temperature stimulus was more reliable across sessions while the opposite was found for other areas. Thus, in this paradigm, no definitive conclusion can be made regarding intersession reliability differences in fMRI activation resulting from stimuli of constant temperature versus stimuli that produce a constant perception. One possible explanation is that variation in perceived pain intensity is not an important contributor to the variability of the BOLD response, a notion that is supported by our finding of only scattered clusters for which pain intensity ratings significantly covaried with the BOLD signal amplitude response to painful stimuli. Another possible explanation for failing to find a consistent difference is that the methodology of the current study (such as variability in the subject pool or choice of stimuli) did not provide the range of data adequate to draw out differences. Thus, the possibility remains that in some populations or contexts, pain-related fMRI activation may be more reliable for stimuli of fixed temperature or fixed perception; studies that further explore this issue may provide valuable information to aid in the design of pain imaging experiments.

4.5. Strengths and limitations

A major strength of this study is that the analytical approach used to assess reliability of pain-related activation was also applied to fMRI data collected from the same subjects during the same sessions as they performed a simple motor task (finger-thumb opposition). Using this approach, intersession reliability of motor-related activation in this study was found to be comparable to previously published results for both spatial overlap and ICC measures (Bennett and Miller, 2010; Gountouna et al., 2010; Havel et al., 2006; Kong et al., 2007; McGregor et al., 2012; Yoo et al., 2005). This supports the validity and appropriateness of the analytical approach used in this study to assess intersession reliability of pain-related fMRI activation.

This study is the first to quantitatively examine intersession reliability of pain-related fMRI activation in the entire cortical pain network, expanding upon the work of Taylor and Davis (2009), who limited their analysis to somatosensory processing regions. Furthermore, this study is the first to examine reliability of activation evoked by a ramp-and-hold contact heat stimulus paradigm, which is used in many pain imaging paradigms, including those assessing changes associated with pain-reducing manipulations. While the sample size of 14 subjects in this study is modest, it is the largest used to evaluate intersession reliability of pain-related fMRI activation and is greater than the average number of 11 subjects across all test–retest fMRI activation studies (Bennett and Miller, 2010). Furthermore, the diversity of our subjects in terms of age and gender, which likely added to inter-individual variability of brain responses, can be viewed as a strength, as the results should be more generalizable than those obtained from a more homogeneous subject population such as undergraduate students, as utilized in many test–retest fMRI studies (Bennett and Miller, 2010).

Habituation is a possible explanation for finding poor intersession pain-related fMRI reliability in some brain areas. However, consistent with published literature for the type of ramp-and-hold contact heat

stimulus paradigm used in this study (Quiton and Greenspan, 2008), pain intensity ratings did not change across the three sessions and the temperatures required to produce subject-specific ratings of 50 on a 0–100 scale did not change over the course of the experiments. Furthermore, the data show no systematic decreases in either number of significant voxels or signal amplitude changes across the three sessions. Bingel et al. (2008) required daily application of 60 painful heat stimuli to the forearm for 8 days before observing perceptual habituation and decreases in pain-related fMRI responses in the brain. Thus, it is unlikely that habituation is a major contributor to the poor intersession reliability observed for some brain areas in this study.

Intersession reliability may vary based on the number of trials of painful stimuli. Conclusions from this study, in which 12 trials of each temperature were applied in each session, may not be generalizable to studies where the number of trials is significantly different. Increasing the number of stimuli has the potential to increase statistical power and intersession reliability; however, it also might introduce dynamic changes in the BOLD response that would add variability and thereby decrease intersession reliability. A separate study is needed to determine how reliability varies as a function of stimulus number.

4.6. Conclusions

Overall, in this paradigm, pain-related BOLD fMRI responses showed fair to moderate test–retest reliability in brain regions that are part of the classical pain network. A review of over 63 studies of fMRI test–retest reliability for various tasks, designs, and test–retest intervals reported the average ICC value was 0.5 (Bennett and Miller, 2010). The ICCs calculated in this study for most brain regions were greater than 0.5, suggesting that pain-related fMRI activation has better than average intersession reliability. The finding that some brain regions showed stronger test–retest reliability than others may provide useful information to guide longitudinal pain studies. In addition, the study findings led to the following recommendations for test–retest reliability analyses: (1) Spatial extent and localization of activation do not appear to be useful measures of fMRI response reliability at the group level. (2) The use of ICC values alone as a measure of reliability may not be sufficient, as the underlying variance structure of a dataset may result in erroneously high ICC values. Methods to eliminate erroneously high ICC values (such as the filtering step used in this study) should be incorporated into any analysis of this type.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2014.07.005.

Acknowledgements

The authors gratefully acknowledge NIH for funding (R03-AG022223/R01-NS39337). The authors of this manuscript do not have any financial or other relationships that might lead to a conflict of interest.

References

Atri, A., O'Brien, J.L., Sreenivasan, A., Rastegar, S., Salisbury, S., DeLuca, A.N., O'Keefe, K.M., LaViolette, P.S., Rentz, D.M., Locascio, J.J., Sperling, R.A., 2011. Test–retest reliability of memory task fMRI in Alzheimer's disease clinical trials. Archives of Neurology 68, 599–606. http://dx.doi.org/10.1001/archneurol.2011.9421555634.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Annals of the New York Academy of Sciences 1191, 133–155. http://dx.doi.org/10.1111/j.1749-6632.2010.05446.x20392279.

Berthier, M., Starkstein, S., Leiguarda, R., 1988. Asymbolia for pain: A sensory-limbic disconnection syndrome. Annals of Neurology 24, 41–49. http://dx.doi.org/10.1002/ana.4102401093415199.

Bingel, U., Herken, W., Teutsch, S., May, A., 2008. Letter to the editor: Habituation to painful stimulation involves the antinociceptive system — A 1-year follow-up of 10 participants. Pain 140, 393–394. http://dx.doi.org/10.1016/j.pain.2008.09.03018952372.

Bushnell, M.C., Duncan, G.H., Hofbauer, R.K., Ha, B., Chen, J.-I., Carrier, B., 1999. Pain perception: Is there a role for primary somatosensory cortex? Proceedings of the National Academy of Sciences of the United States of America 96, 7705–770910393884.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45, 758–768. http://dx.doi.org/10.1016/j.neuroimage.2008.12.03519166942.

Chen, E.E., Small, S.L., 2007. Test–retest reliability in fMRI of language: Group and task effects. Brain and Language 102, 176–185. http://dx.doi.org/10.1016/j.bandl.2006.04.01516753206.

Cox, R.W., 1996. AFNI: Software for analysis and visualization of functional magnetic resonance Neuroimages. Computers and Biomedical Research, an International Journal 29, 162–1738812068.

Craig, A.D., 2002. How do you feel? Interoception: The sense of the physiological condition of the body. Nature Reviews. Neuroscience 3, 655–666. http://dx.doi.org/10.1038/nrn89412154366.

Craig, A.D., 2003. Interoception: The sense of the physiological condition of the body. Current Opinion in Neurobiology 13, 500–50512965300.

Craig, A.D., 2009. How do you feel — Now? The anterior insula and human awareness. Nature Reviews. Neuroscience 10, 59–70. http://dx.doi.org/10.1038/nrn255519096369.

Craig, A.D., 2011. Significance of the insula for the evolution of human awareness of feelings from the body. Annals of the New York Academy of Sciences 1225, 72–82. http://dx.doi.org/10.1111/j.1749-6632.2011.05990.x21534994.

Donner, A., Zou, G., 2002. Testing equality of dependent intraclass correlation coefficients. Journal of the Royal Statistical Society, Series D (The Statistician) 51, 367–379.

Duerden, E.G., Albanese, M.-C., 2013. Localization of pain-related brain activation: A meta-analysis of neuroimaging data. Human Brain Mapping 34, 109–149. http://dx.doi.org/10.1002/hbm.2141622131304.

Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. NeuroImage 50, 1168–1176. http://dx.doi.org/10.1016/j.neuroimage.2010.01.03620083206.

Freyer, T., Valerius, G., Kuelz, A.-K., Speck, O., Glauche, V., Hull, M., Voderholzer, U., 2009. Test–retest reliability of event-related functional MRI in a probabilistic reversal learning task. Psychiatry Research 174, 40–46. http://dx.doi.org/10.1016/j.pscychresns.2009.03.00319783412.

Friedman, D.P., Murray, E.A., O'Neill, J.B., Mishkin, M., 1986. Cortical connections of the somatosensory fields of the lateral sulcus of macaques: Evidence for a corticolimbic pathway for touch. Journal of Comparative Neurology 252, 323–347. http://dx.doi.org/10.1002/cne.9025203043793980.

Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013. A test–retest fMRI dataset for motor, language, and spatial attention functions. GigaScience 2, 6. http://dx.doi.org/10.1186/2047-217X-2-623628139.

Gountouna, V.-E., Job, D.E., McIntosh, A.M., Moorhead, T.W.J., Lymer, G.K.L., Whalley, H.C., Hall, J., Waiter, G.D., Brennan, D., McGonigle, D.J., Ahearn, T.S., Cavanagh, J., Condon, B., Hadley, D.M., Marshall, I., Murray, A.D., Steele, J.D., Wardlaw, J.M., Lawrie, S.M., 2010. Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. NeuroImage 49, 552–560. http://dx.doi.org/10.1016/j.neuroimage.2009.07.02619631757.

Havel, P., Braun, B., Rau, S., Tonn, J.-C., Fesl, G., Brückmann, H., Ilmberger, J., 2006. Reproducibility of activation in four motor paradigms. An fMRI study. Journal of Neurology 253, 471–476. http://dx.doi.org/10.1007/s00415-005-0028-416283098.

Heimer, L., Van Hoesen, G.W., 2006. The limbic lobe and its output channels: Implications for emotional functions and adaptive behavior. Neuroscience and Biobehavioral Reviews 30, 126–147. http://dx.doi.org/10.1016/j.neubiorev.2005.06.00616183121.

Kiehl, K.A., Liddle, P.F., 2003. Reproducibility of the hemodynamic response to auditory oddball stimuli: A six-week test–retest study. Human Brain Mapping 18, 42–52. http://dx.doi.org/10.1002/hbm.1007412454911.

Kong, J., Gollub, R.L., Webb, J.M., Kong, J.-T., Vangel, M.G., Kwong, K., 2007. Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. NeuroImage 34, 1171–1181. http://dx.doi.org/10.1016/j.neuroimage.2006.10.01917157035.

Maitra, R., Roys, S.R., Gullapalli, R.P., 2002. Test–retest reliability estimation of functional MRI data. Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 48, 62–70. http://dx.doi.org/10.1002/mrm.1019112111932.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46.

McGregor, K.M., Carpenter, H., Kleim, E., Sudhyadhom, A., White, K.D., Butler, A.J., Kleim, J., Crosson, B., 2012. Motor map reliability and aging: A TMS/fMRI study. Experimental Brain Research 219, 97–106. http://dx.doi.org/10.1007/s00221-012-3070-322466408.

Moulton, E.A., Keaser, M.L., Gullapalli, R.P., Greenspan, J.D., 2005. Regional intensive and temporal patterns of functional MRI activation distinguishing noxious and innocuous contact heat. Journal of Neurophysiology 93, 2183–2193. http://dx.doi.org/10.1152/jn.01025.200415601733.

Quiton, R.L., Greenspan, J.D., 2008. Across- and within-session variability of ratings of painful contact heat stimuli. Pain 137, 245–256. http://dx.doi.org/10.1016/j.pain.2007.08.03417942227.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magnetic Resonance Imaging 16, 105–1139508267.

Servos, P., Zacks, J., Rumelhart, D.E., Glover, G.H., 1998. Somatotopy of the human arm using fMRI. Neuroreport 9, 605–6099559924.

Shrout, P.E., 1998. Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research 7, 301–3179803527.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86, 420–42818839484.

Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statisticsfifth edition. Pearson/Allyn & Bacon, Boston, MA.

Tailairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human BrainThieme Medical Publishers, New York.
Taylor, K.S., Davis, K.D., 2009. Stability of tactile- and pain-related fMRI brain activations: An examination of threshold-dependent and threshold-independent methods. Human Brain Mapping 30, 1947–1962. http://dx.doi.org/10.1002/hbm.2064118711711.
Van den Bosch, G.E., van Hemmen, J., White, T., Tibboel, D., Peters, J.W.B., van der Geest, J.N., 2013. Standard and individually determined thermal pain stimuli induce similar brain activations. European Journal of Pain (London, England) 17, 1307–1315. http://dx.doi.org/10.1002/j.1532-2149.2013.00311.x23529976.
Van Westen, D., Fransson, P., Olsrud, J., Rosén, B., Lundborg, G., Larsson, E.-M., 2004. Fingersomatotopy in area 3b: An fMRI-study. BMC Neuroscience 5, 28. http://dx.doi.org/10.1186/1471-2202-5-2815320953.
Vogt, B.A., 2005. Pain and emotion interactions in subregions of the cingulate gyrus. Nature Reviews. Neuroscience 6, 533–544. http://dx.doi.org/10.1038/nrn170415995724.
Winer, B.J., 1971. Statistical Principles in Experimental Designsecond edition. McGraw-Hill, New York.
Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nature Methods 8, 665–670. http://dx.doi.org/10.1038/nmeth.163521706013.
Yoo, S.-S., Wei, X., Dickey, C.C., Guttmann, C.R.G., Panych, L.P., 2005. Long-term reproducibility analysis of fMRI using hand motor task. International Journal of Neuroscience 115, 55–7715768852.