◆ Human Brain Mapping 36:2558–2579 (2015) ◆

Reliability of an fMRI Paradigm for Emotional Processing in a Multisite Longitudinal Study

**Dylan G. Gee,[1] Sarah C. McEwen,[1] Jennifer K. Forsyth,[1] Kristen M. Haut,[2] Carrie E. Bearden,[1] Jean Addington,[3] Bradley Goodyear,[3] Kristin S. Cadenhead,[4] Heline Mirzakhanian,[4] Barbara A. Cornblatt,[5] Doreen Olvet,[5] Daniel H. Mathalon,[6] Thomas H. McGlashan,[7] Diana O. Perkins,[8] Aysenil Belger,[8] Larry J. Seidman,[9] Heidi Thermenos,[9] Ming T. Tsuang,[4] Theo G.M. van Erp,[10] Elaine F. Walker,[11] Stephan Hamann,[11] Scott W. Woods,[7] Todd Constable,[7] and Tyrone D. Cannon[2,7]*

[1]*Departments of Psychology and Psychiatry, University of California, Los Angeles, California*
[2]*Department of Psychology, Yale University, New Haven, Connecticut*
[3]*Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada*
[4]*Department of Psychiatry, University of California, San Diego, La Jolla, California*
[5]*Department of Psychiatry Research, Zucker Hillside Hospital, Glen Oaks, New York*
[6]*Department of Psychiatry, University of California, San Francisco, California*
[7]*Department of Psychiatry, Yale University, New Haven, Connecticut*
[8]*Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina*
[9]*Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts*
[10]*Department of Psychiatry and Human Behavior, University of California, Irvine, California*
[11]*Department of Psychology, Emory University, Atlanta, Georgia*

Abstract: Multisite neuroimaging studies can facilitate the investigation of brain-related changes in many contexts, including patient groups that are relatively rare in the general population. Though multisite studies have characterized the reliability of brain activation during working memory and motor functional magnetic resonance imaging tasks, emotion processing tasks, pertinent to many clinical populations, remain less explored. A traveling participants study was conducted with eight healthy volunteers scanned twice on consecutive days at each of the eight North American Longitudinal Prodrome Study sites. Tests derived from generalizability theory showed excellent reliability in the amygdala ($E_\rho^2 = 0.82$), inferior frontal gyrus (IFG; $E_\rho^2 = 0.83$), anterior cingulate cortex (ACC; $E_\rho^2 = 0.76$), insula ($E_\rho^2 = 0.85$), and fusiform gyrus ($E_\rho^2 = 0.91$) for maximum activation and fair to excellent reliabil-

Additional Supporting Information may be found in the online version of this article.

Contract grant sponsor: National Institute of Mental Health at the National Institutes of Health (Collaborative U01 award); Contract grant number: MH081902 (to T.D.C.); MH081988 (to E.W.); MH081928 (to L.J.S.); MH081984 (to J.A.); and MH066160 (to S.W.W.); Contract grant sponsor: National Science Foundation Graduate Research Fellowship (to D.G.G.).

*Correspondence to: Tyrone D. Cannon; Yale University, P.O. Box 208205, 2 Hillhouse Avenue, New Haven, CT 06520. E-mail: tyrone.cannon@yale.edu

The authors declare there are no conflicts of interest to report.

Received for publication 28 October 2014; Revised 3 March 2015; Accepted 6 March 2015.

DOI: 10.1002/hbm.22791

Published online 28 March 2015 in Wiley Online Library (wileyonlinelibrary.com).

ity in the amygdala ($E^2_\rho = 0.44$), IFG ($E^2_\rho = 0.48$), ACC ($E^2_\rho = 0.55$), insula ($E^2_\rho = 0.42$), and fusiform gyrus ($E^2_\rho = 0.83$) for mean activation across sites and test days. For the amygdala, habituation ($E^2_\rho = 0.71$) was more stable than mean activation. In a second investigation, data from 111 healthy individuals across sites were aggregated in a voxelwise, quantitative meta-analysis. When compared with a mixed effects model controlling for site, both approaches identified robust activation in regions consistent with expected results based on prior single-site research. Overall, regions central to emotion processing showed strong reliability in the traveling participants study and robust activation in the aggregation study. These results support the reliability of blood oxygen level-dependent signal in emotion processing areas across different sites and scanners and may inform future efforts to increase efficiency and enhance knowledge of rare conditions in the population through multisite neuroimaging paradigms. *Hum Brain Mapp 36:2558–2579, 2015.*

Key words: fMRI; reliability; multisite; emotion; meta-analysis; amygdale

INTRODUCTION

Multisite studies have increasingly facilitated the investigation of brain-related phenomena through neuroimaging. Multisite investigations provide an efficient way to recruit a large sample of participants, enhancing statistical power and the generalizability of results, and have thus contributed critical insight into conditions that are relatively rare in the general population [Addington et al., 2007; Beckett et al., 2010; Cannon et al., 2008; Ford et al., 2009; Kim et al., 2009; Lieberman et al., 2005; Potkin et al., 2009; You et al., 2011]. These advantages make multisite studies a key to future scientific discovery [Glover et al., 2012; Meyer-Lindenberg, 2012]. However, aggregating neuroimaging data across different sites and scanners presents challenges due to numerous possible sources of variability other than individual participant effects [Friedman and Glover, 2006; Ojemann et al., 1998; Pearlson, 2009; Van Horn and Toga, 2009; Voyvodic, 2006; Zou et al., 2005]. Given potential site-related variance and important differences in task design and reproducibility, conducting multisite studies necessitates thorough evaluations of multisite effects and reliability for each functional magnetic resonance imaging (fMRI) task used [Bennett and Miller, 2010; Brown et al., 2011; Friedman et al., 2008; Glover et al., 2012; Pearlson, 2009; Plichta et al., 2012].

Evaluating an fMRI paradigm for multisite implementation involves quantifying the variability in activation related to site effects, such as different scanners and acquisition protocols, and comparing it to the variability introduced by other factors such as participant differences and imaging noise. Traveling participant designs, in which participants are scanned at each site of a multisite study, uniquely allow for the comparison of variance in blood oxygen level-dependent (BOLD) signal due to person- versus site-related factors [e.g., Brown et al., 2011; Gountouna et al., 2010; Gradin et al., 2010; Yendiki et al., 2010]. If activation measures show greater variation related to person than site, person-related effects are likely to generalize across sites and would support the aggregation of data across sites. Alternatively, the generalizability of data across sites would be questionable in the case of greater variation due to site-related differences [Gradin et al., 2010]. Moreover, evaluating an fMRI paradigm at the multisite level necessitates testing the extent to which the aggregation of data across sites produces activation effects that are consistent with hypothesized task-related neural processes and previous findings in single-site investigations. Thus, examining and comparing statistical methods of aggregating data across sites is critical to ensuring that methods for pooling data are both valid and maximize the potential advantages offered by multisite studies.

Variance component estimates can be used to compute reliability coefficients that provide summary statistics for the consistency of measurement across multiple assessments. Reliability generally refers to consistency in the ranking of persons on a given measure over multiple assessments, and reliability coefficients can be calculated to assess relative or absolute reliability, depending on the nature of the decisions to be made from the measurements. Relative decisions are based on an individual's measurement relative to the measurements obtained from others (e.g., norm-referenced interpretations of measurements), whereas absolute decisions are based on the absolute level of an individual's measurement independent of the measurements obtained from others [Shavelson and Webb, 1991]. The distinction mainly concerns whether the main effects of a facet of observation (such as test item, measurement occasion, or in the current context, MRI scanner) are considered to contribute to measurement error and included in the error term of the reliability coefficient. In the case of relative decisions, they are not included, whereas they are included in the case of absolute decisions. Measures of relative reliability include the generalizability coefficient (G-coefficient) of generalizability theory [Brennan, 2001; Shavelson and Webb, 1991], the intraclass correlation (ICC; Type 3,1) statistic of Shrout and Fleiss [1979], and the Pearson correlation coefficient. Measures of absolute reliability include the absolute level ICC (Type 2,1) [Shrout and Fleiss, 1979] or the dependability

coefficient (D-coefficient) of generalizability theory [Shavelson and Webb, 1991]. Given that the primary aim of many fMRI studies involves describing group differences between task contrasts or describing correlations between task contrasts and other variables of interest [i.e., relative decisions; Barch and Mathalon, 2011], assessing relative agreement across scanning sites may be appropriate in the context of multisite fMRI studies. However, in cases where the absolute value of activation will be utilized for interpretation or in multisite studies in which scanning site is not independent of other factors, assessing the absolute agreement of fMRI measurement across sites may also be valuable [Brown et al., 2011]. For example, if there are significant differences in the ratio of case versus control participants across sites in a multisite study, adjusting for site in the analysis may not be sufficient to eliminate all site effects. In such circumstances, assessment of reliability at an absolute level would inform the extent to which data are interchangeable across sites and thus the extent to which merging fMRI data across sites is valid [Friedman et al., 2008]. The most appropriate reliability measure, therefore, depends on study design and the research question at hand.

To date, several studies have demonstrated that measures of BOLD signal are replicable across sites with the same scanner models for cognitive task paradigms such as working memory [Bernal-Casas et al., 2013; Brandt et al., 2013; Brown et al., 2011; Casey et al., 1998; Costafreda et al., 2007; Forsyth et al., 2014; Friedman et al., 2008; Yendiki et al., 2010], as well as motor tasks [Costafreda et al., 2007; Gountouna et al., 2010; Sutton et al., 2008; Wurnig et al., 2013] and resting-state scans [Huang et al., 2012]. Specifically, results showed that across-participant variability was greater than across-site variability. In addition, prior work suggests that aggregating data across sites necessitates only a modest increase in sample size to compensate for decreases in power due to the addition of site-related variance [Suckling et al., 2008]. However, the development and evaluation of analytic approaches for aggregating multisite data is greatly needed to inform future multisite investigations [Costafreda, 2009]. Finally, studies of multisite reliability must expand beyond motor and working memory paradigms to investigate other domains that are highly relevant for clinical populations, such as emotion processing. A study of an implicit sad affect paradigm characterized reliability across two sites [Suckling et al., 2008], providing important initial information about the potential of multisite affective paradigms; however, multisite reliability data and analytic approaches are needed for paradigms that actively probe emotional processes such as emotion perception, identification, and regulation.

This study aimed to characterize the reliability of activation during an emotional faces fMRI task as part of a large ongoing, multisite study of neurobiological risk factors for psychosis onset. Emotion processing represents a core domain of impairment in schizophrenia [Kring and Moran, 2008; Mueser et al., 1996]; however, the extent to which deficits in emotion processing are present prior to the onset of psychosis and the role that they might play in its development remain unclear. Thus, a primary aim of the neuroimaging component of the North American Prodrome Longitudinal Study [NAPLS; Addington et al., 2007] was to examine neural circuitry associated with emotion processing and developmental trajectories of these regions in clinical high risk (CHR) patients, as compared with healthy controls. The NAPLS study used a well-established emotional faces fMRI task [Fakra et al., 2008; Gee et al., 2012; Hariri et al., 2000; Lieberman et al., 2007], which was selected for several empirical reasons. The task has been widely used across the emotion regulation literature, and thus a wealth of findings have characterized the task and its neural correlates among healthy controls [Hariri et al., 2000; Lieberman et al., 2007, 2011; Tabibnia et al., 2008]. Specifically, this paradigm has been shown to robustly activate the amygdala and prefrontal cortex, regions that are central to emotional processing, in healthy controls [Lieberman et al., 2007]. In particular, increased inferior frontal gyrus (IFG) activation is observed when participants label the emotional expressions portrayed on faces, and increased amygdala activation is observed when participants match faces based on emotional expression. Moreover, patients with schizophrenia show impairment and alterations in neural activation during the same emotional faces task [Fakra et al., 2008], making it a prime candidate for the study of early disruptions in relevant neural circuitry among CHR patients. Finally, because the paradigm is in widespread use at the single-site level, substantial data exist with which to compare the present multisite results, and it is likely that research on its reliability would be relevant to the broader fields of social neuroscience and psychopathology.

To examine the reliability of fMRI activation during the emotional faces task across the eight NAPLS sites and to establish valid statistical methods for aggregating fMRI data across sites, we present data from two study samples here. For the first study sample, we used a traveling participant study design and generalizability theory to characterize the proportion of variance in BOLD signal attributable to person- versus site-related factors and to assess the reliability of the person effect across scanning sites and days at both a relative and an absolute level. We predicted that variance in activation due to person-related factors would be greater than variance due to site-related factors and that the person effect would be reliable across sites for the amygdala and IFG. Given recent evidence that habituation has more stable test-retest reliability than amplitude of activation for the amygdala (Plichta et al., 2014), we also examined amygdala habituation. In the second study sample, fMRI data for all healthy individuals who had been recruited as control participants in the NAPLS study (for comparison to the CHR sample) were aggregated across sites using two statistical methods. We

TABLE I. Participant characteristics for traveling participants study, aggregation study, and aggregation study by site

	Traveling participants study	Aggregation study	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
n	8	111	11	15	16	18	18	7	11	15
Age	26.9 (4.3)	21.04 (4.7)	19.8 (3.2)	22.9 (5.5)	19.9 (3.9)	17.6 (2.4)	20.8 (3.0)	21.9 (5.2)	24.6 (5.9)	22.7 (5.1)
Range	20–31	12–33	16–26	13–31	14–28	12–23	14–25	14–29	15–33	14–30
Sex	4F/4M	48F/63M	3F/8M	6F/9M	8F/8M	8F/10M	6F/12M	2F/5M	5F/6M	10F/5M

assessed similarities and differences in results from these two approaches and compared the present multisite findings to single-site studies using the same emotional faces task.

MATERIALS AND METHODS

Participants

This work comprised two studies of the multisite NAPLS investigation with two separate participant samples. Study 1 ("traveling participants study") examined the reliability of activation across sites, participants, and testing days. Eight healthy participants were recruited (one from each of the eight sites) for the traveling participants component of the larger NAPLS study. Due to the significant travel requirement, only participants 18 years and older were recruited for that component. The traveling participants ranged in age from 20 to 31 years old (mean = 26.9, S.D. = 4.3). These eight participants traveled to all eight sites and were scanned twice on consecutive days at each site, yielding a set of 128 scans (eight participants × two scans × eight sites). All traveling participants completed all scans within four months (May through August of 2011). The order of visits to the eight sites was counterbalanced across participants.

Study 2 ("aggregation study") examined fMRI activation when data from the eight NAPLS sites were aggregated. For the aggregation study, a total of 111 unique healthy controls between the ages of 12 and 33 years old (mean = 21.0, S.D. = 4.7) were scanned at the NAPLS site at which they were recruited (see Table I for sample sizes at each site).

For both samples, participants were excluded if they met DSM-IV criteria for a psychiatric disorder (as assessed with the Structured Clinical Interview for DSM-IV-TR; First et al., [2002] or Kiddie-Schedule for Affective Disorders and Schizophrenia; Kaufman et al., [1997]), had a first-degree relative with a current or past psychotic disorder, met prodromal criteria (as assessed by the Structured Interview for Prodromal Syndromes) [McGlashan et al., 2001], met criteria for substance dependence (in the past 6 months), had a neurological disorder, or had a Full Scale IQ <70 (as measured by the Wechsler Abbreviated Scale of Intelligence) [Wechsler, 1999].

All participants provided informed consent or assent for the study and were compensated for their participation. Participants were recruited from the community via advertising. Parental informed consent for minors was also obtained. The protocol was approved by Institutional Review Boards at the sites participating in the NAPLS, from which participants were drawn (Emory University, Harvard University, University of Calgary, University of California Los Angeles (UCLA), University of California San Diego, University of North Carolina (UNC), Yale University, Zucker Hillside Hospital).

Task Design

The experimental paradigm consisted of an emotional faces task [Hariri et al., 2000; Lieberman et al., 2007]. Participants viewed target faces or shapes while performing one of five tasks in each block of 10 trials (each trial was 5 s; i.e., 50-s blocks in randomized order). Emotion labeling involved choosing which of two labels (e.g., "angry," "happy," "scared," "surprised") described a target face. Gender labeling involved selecting the gender-appropriate name for a target face. Emotion matching involved choosing which of two faces displayed the same emotion as a target face. Gender matching involved selecting which of two faces was the same gender as a target face. Shape matching involved selecting which of two shapes was the same as a target shape. Directions for the task were reviewed prior to each scanning session.

The facial stimuli were chosen from a standardized set of images [Tottenham et al., 2009]. Half of the target faces in each condition were female and half were male. The target face depicted a negative emotional expression (i.e., fear or anger) in 80% of the trials comprising each condition. In the other 20% of trials, the target face consisted of a happy or surprised face. The emotion labels and gender names were matched on a number of dimensions (same number of words, word length; for each emotion label, there was a name that began with the same letter in the gender label condition). Correct responses were on the left 50% of the time and on the right 50% of the time. A given

identity did not appear more than one time in any given block.

Each block began with a 3-s instruction cue to indicate the task condition, followed by 10 trials of that task, randomly selected from a pool of trials. Blocks were separated by a 10-s fixation crosshair. Each of the five conditions appeared once within each run. Each condition was represented in a separate block, such that five blocks comprised each run. Within each run, the order of the conditions was randomized. Participants completed two functional runs (i.e., two blocks, or 20 trials, of each condition). Responses were registered using a button box, and participants were told to respond as soon as they were sure of the correct answer. The stimuli remained on the screen for the entire 5-s duration of each trial. To minimize practice effects, four different versions of the task were used and counterbalanced across scans. In the four parallel versions, we systematically varied which identity appeared in which condition and which identity was paired with each facial expression.

Emotion labeling is considered to represent a form of incidental emotion regulation [Lieberman et al., 2007] and has been associated with increased activation in the IFG and dampening of amygdala activation, as well as decreases in negative emotion and reduced physiological responsivity [Lieberman et al., 2011; Tabibnia et al., 2008]. By contrast, emotion matching has been associated with increased amygdala activation, relative to emotion labeling. For these reasons, we focused on emotion labeling for the IFG and emotion matching for the amygdala to probe amygdala-prefrontal circuitry in this study. In addition, we examined several regions that are commonly activated during emotion paradigms [Fusar-Poli et al., 2009; Kober et al., 2008; Phan et al., 2002] to enhance generalizability to other studies of emotion processing. Activation in the anterior cingulate cortex (ACC), insula, and fusiform gyrus was measured during emotion matching. Emotion matching was contrasted with the control condition of shape matching to isolate the processing of emotional faces while controlling for the process of matching. Emotion labeling was contrasted with the control condition of gender labeling to isolate emotion processing while controlling for the process of labeling. In addition, emotion labeling and emotion matching were directly contrasted. Emotion labeling and emotion matching were also compared with implicit baseline (consisting of unmodeled fixation events during the intertrial intervals). We selected these contrasts based on prior work with this task [Fakra et al., 2008; Hariri et al., 2000; Lieberman et al., 2007; Plichta et al., 2014].

Behavioral Data Analysis

For each participant, mean accuracy (percentage correct) and mean reaction time (RT) were calculated for each individual condition (i.e., emotion labeling, emotion matching, gender labeling, gender matching, shape matching) and across the entire task (mean performance across all five conditions). An analysis of variance (ANOVA) was used to examine potential differences in task performance (mean accuracy, mean RT) between sites.

Data Acquisition

Scanning was performed on Siemens Trio 3T scanners at UCLA, Emory, Harvard, UNC, and Yale, GE 3T HDx scanners at Zucker Hillside Hospital and UCSD, and a GE 3T Discovery scanner at Calgary. Due to longevity of the study and scanner repairs, data for eight participants ($n = 2$ at UCLA, $n = 1$ at Emory, $n = 2$ at Harvard, $n = 1$ at UCSD, $n = 2$ at Calgary) from the larger sample of healthy controls were collected on alternate scanners, and one traveling participant received both scans at Harvard on an alternate scanner. To facilitate reliability analyses, data from these scans were excluded from analysis for the current study. All Siemens sites used a 12-channel head coil and all GE sites used an 8-channel head coil. Head movements were restricted with foam padding. Anatomical reference scans were acquired first and used to configure slice alignment. A T2-weighted image (0.9-mm in-plane resolution) was acquired using a set of high-resolution echo planar (EPI) localizers (Siemens: TR/TE 6,310/67 ms, 30 4-mm slices with 1-mm gap, 220-mm FOV; GE: TR/TE 6,000/120 ms, 30 4-mm slices with 1-mm gap, 220-mm FOV). Functional scans matched the AC-PC aligned T2 image and utilized an EPI sequence (TR/TE 2,500/30 ms, 77 degree flip angle, 30 4-mm slices). The task consisted of two functional runs, each of 129 volumes. Stimuli were presented using a PC running E-Prime (Psychology Software Tools), and participants viewed the stimuli via LCD goggles (Resonance Technologies, Inc.).

fMRI Data Analysis

Functional image analysis was performed using FSL (FMRIB's Software Library v. 4.0) [Smith et al., 2004]. Motion in EPI data was corrected using a six-parameter, rigid-body three-dimensional coregistration (FLIRT; FMRIB's Linear Image Registration Tool), which registered each BOLD image to the middle data point in the timeseries. Data were registered for each participant (EPI to participant's T2-weighted structural image, then T2 to standard space brain) [Jenkinson and Smith, 2001; Jenkinson et al., 2002]. Data were spatially smoothed with a 5-mm (FWHM; full width at half maximum) Gaussian kernel and filtered with a nonlinear high-pass filter (120 s cutoff). Individual participant analyses used FEAT (FMRI Expert Analysis Tool).

Timeseries statistical analysis on each participant was performed using FILM (FMRIB's Improved Linear Model) with local autocorrelation correction [Woolrich et al., 2001]. A univariate general linear model (GLM) was applied on a voxel-by-voxel basis such that each voxel's

timeseries was individually fitted to the resulting model, with local autocorrelation correction applied within tissue type to improve temporal smoothness estimation [Smith et al., 2004; Woolrich et al., 2001]. Each voxel's goodness-of-fit to the model was estimated; resulting parameter estimates indicated the degree to which signal change could be explained by each model. Each condition was modeled separately, with each correct trial modeled in its entirety in a block design fashion. Motion parameters were entered as covariates to correct for head motion artifacts; volumes with motion exceeding 3 mm were excluded from analysis. For the traveling subjects study, 3 of the 128 scans (from three different participants) exceeded the motion threshold with 2, 4, and 45 of the 129 volumes excluded, respectively. Thus, for the entire set of the traveling subjects scans, the average % of excluded volumes was 0.3% (median = 0%, mode = 0%). For the aggregation study, 9 of the 111 participants exceeded the motion threshold, with an average of 4.7 (range = 2–7) of the 129 volumes excluded (mean % volumes censored for those nine participants = 3.6%). Thus, for the entire sample of the 111 participants, the average percentage of excluded volumes was 0.2% (median = 0%, mode = 0%). Resulting contrast images were entered into second-level analyses using a fixed effects model to combine functional runs for each participant and to allow for inferences at the group level. To correct for multiple comparisons, resulting Z-statistic images were thresholded using clusters determined by $Z > 2.3$ and a corrected cluster significance threshold of $P = 0.05$ (Forman et al., 1995; Friston, 1994; Worsley et al., 1992). Cluster *P*-values were determined using spatial smoothness estimation in FEAT (Forman et al., 1995; Friston, 1994; Jenkinson and Smith, 2001).

To check the quality of data and minimize variability between sites, a quality assurance protocol was implemented across sites. Functional data were checked for motion, artifacts, and the quality of skull stripping implemented in FSL, and data diagnostics were checked for each participant.

Regions of Interest

Regions of interest (ROI) were selected based on the primary findings of prior work [Hariri et al., 2000; Lieberman et al., 2007], which has consistently observed amygdala and IFG activation during the emotional faces task. Two types of ROI masks were examined: (1) anatomically defined masks and (2) combined functional and anatomical masks. Anatomically defined masks were created using the Harvard-Oxford Structural Atlas [Kennedy et al., 1998; Makris et al., 1999] for the amygdala, inferior frontal gyrus, ACC, insula, and fusiform gyrus (labeled as temporal occipital fusiform cortex). Voxels with atlas-derived values corresponding to a probability greater than 25% of belonging to the given region were included. Combined anatomical and functional masks were created to probe specific regions of the anatomical structures that were activated during the present task. Specifically, a group activation map for all traveling participant scans (i.e., 128 scans) was used to provide a functional map for a given ROI. Based on prior work, the contrast of emotion labeling > baseline was used for the IFG and emotion matching > baseline was used for the amygdala. Group-level activation was evident at a cluster-corrected threshold of $z = 2.3$ for all regions, with the exception of the left amygdala ($z = 2.0$) and ACC (thus only an anatomical mask was used for the ACC). A convergence analysis was used to create a mask of the voxels that overlapped between the functional maps and anatomical masks for each of the ROI. See Supporting Information Methods for quality assessment of ROI definition.

FSL's Featquery was used to warp ROIs back into each participant's space by applying the inverse of the transformation matrix used during the initial registration. For each anatomical and combined anatomical/functional ROI, we obtained indices of activation using the contrasts of parameter estimates from both scans in a single scanning session (second-level fixed effects analysis). The motion-corrected, smoothed, and filtered data were probed for mean and maximum percent signal change for each contrast of interest. Amygdala habituation was calculated based on the amplitude difference between the first and second run of the task for each condition (Blackford et al., 2013; Plichta et al., 2014).

STUDY 1: TRAVELING PARTICIPANTS STUDY

Determining Reliability Using G-Theory

Reliability of activation for each ROI was assessed using the generalizability theory (G-theory) framework. G-theory was developed as an extension of classical test theory to recognize and model the multiple sources of measurement error that influence a measure's reliability, or generalizability, and to allow estimation of reliability with respect to only those sources of error relevant to the research questions at hand [Barch and Mathalon, 2011]. Briefly, reliability assessment using G-theory includes a generalizability study (G-study) and a decision study (D-study). The G-study extends earlier ANOVA approaches to reliability by partitioning total variance in scores into the variance components associated with: (1) the main effect of person (i.e., the object of measurement); (2) the main effect of each characteristic feature of the measurement situation such as test site, test occasion, or test form, termed "facets" of measurement; and (3) their interactions. The objects of measurement (i.e., persons) are considered to be sampled from a population, and variability among persons is referred to as "universe score variance." A "universe of admissible observations" is thus defined by all possible combinations of all the levels of the facets. G-theory describes the dependability or reliability of generalizations

made from a person's observed score to the score he or she would obtain in the broad universe of admissible observations.

G-theory distinguishes between reliability based on the relative standing of persons versus those based on the absolute value of a score in the subsequent D-study [Di Nocera et al., 2001]. For relative decisions, the estimated components of variance from the G-study are used to compute a generalizability coefficient (G-coefficient), which is the ratio of the universe score variance to itself plus relative error variance. As such, the G-coefficient is an intraclass correlation and is analogous to a reliability coefficient in classical test theory. G-coefficients vary between 0 and 1 and describe the reliability of the rank ordering of individuals. The error term (σ^2_{rel}) of the G-coefficient, E^2_ρ, arises from all the non-zero variance components associated with the rank ordering of individuals. Thus, variance components associated with the interaction of person with each facet or combination of facets define the error term. The G-coefficient is expressed as:

$$E^2_\rho = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{rel})};$$

where σ^2_p represents the variance in scores due to person.

For absolute decisions, estimated components of variance from the G-study are used to compute an index of dependability (D-coefficient). The error term (σ^2_{abs}) of the D-coefficient (φ) arises from all the variance components associated with the score aside from the component associated with the object of measurement. The D-coefficient represents the reliability of 1 observed value within the universe of admissible observations and similarly varies from 0 to 1. It is expressed as the following:

$$\varphi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{abs})}$$

For a more detailed discussion of G-theory see Webb and Shavelson [2005].

Statistical Analyses

The G-study was performed using a two-facet Person (8 levels) × Site (8 levels) × Testing Day (2 levels) crossed design. Person represented the object of measurement and was crossed with the site and day facets. Thus, variance components were estimated for the main effects of person (σ^2_p), site (σ^2_s), and day (σ^2_d); the two-way interactions between person and site (σ^2_{ps}), person and day (σ^2_{pd}), and site and day (σ^2_{sd}); and the residual due to the person × site × day interaction and random error ($\sigma^2_{psd,e}$). The design can be summarized as:

$$\sigma^2(X_{psd}) = \sigma^2_p + \sigma^2_s + \sigma^2_d + \sigma^2_{ps} + \sigma^2_{pd} + \sigma^2_{sd} + \sigma^2_{psd,e};$$

where X_{psd} represents the observed activation score for a person (p) at a site (s) on a testing day (d). All facets were specified as random to maximize generalizability of results to all conditions within facets, including those not explicitly included in the current study. The VARCOMP procedure in SAS with the restricted maximum likelihood method specified was used to estimate variance components for mean and maximum percent signal change for each ROI. In addition, the VARCOMP procedure was used to estimate variance components for accuracy and mean RT for each condition. Any observations excluded from analyses were treated as missing data by the VARCOMP procedure; variance components were estimated on remaining observations.

In the D-study, we investigated the extent to which both the relative ranking of persons and the absolute value of activation for each person was reliable, or generalizable, across scanning sites and test days. Estimated variance components from the G-study were, therefore, used to calculate G-coefficients and D-coefficients that describe the relative and absolute reliability, respectively, of the person effect for activation in each ROI across scanning sites and testing occasions. Reliability coefficients were interpreted using Cicchetti and Sparrow's [1981] definition for judging the clinical significance of ICC values: <0.40 poor; 0.40–0.59 fair; 0.60–0.74 good; >0.74 excellent. G-Coefficients were calculated according to the following equation:

$$E^2_\rho = \frac{\sigma^2_p}{\left(\sigma^2_p + \frac{\sigma^2_{ps}}{n'_s} + \frac{\sigma^2_{pd}}{n'_d} + \frac{\sigma^2_{psd,e}}{n'_s n'_d}\right)}$$

D-coefficients were calculated according to the following equation:

$$\varphi = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_s}{n'_s} + \frac{\sigma^2_d}{n'_d} + \frac{\sigma^2_{ps}}{n'_s} + \frac{\sigma^2_{pd}}{n'_d} + \frac{\sigma^2_{sd}}{n'_s n'_d} + \frac{\sigma^2_{psd,e}}{n'_s n'_d}}$$

In addition to calculating G-coefficients and D-coefficients of reliability, we also conducted an ANOVA for each ROI and contrast to test whether the effects of site, day, or a site-by-day interaction on percent signal change were significant.

STUDY 2: AGGREGATION STUDY

Aggregation of Multisite Data: Hierarchical Model for Image-Based Meta-Analysis

The image-based meta-analysis approach (IBMA) was selected based on prior research comparing different image-based strategies for pooling data across studies [Salimi-Khorshidi et al., 2009]. For each individual site, a mixed effects group-level analysis was performed using FLAME (FMRIB's Local Analysis of Mixed Effects) [Behrens et al., 2003; Smith et al., 2004] with each participant's data, including parameter and variance estimates for each contrast from the lower-level analysis. This interparticipant analysis for each site constituted the third level of fMRI

analysis (following the first-level intra-participant modeling of each participant's fMRI time series data and the second-level fixed effects analysis to combine data from both runs for each participant). The GLM for each site controlled for age and sex. To correct for multiple comparisons, resulting *Z*-statistic images were thresholded using clusters determined by $Z > 2.3$ and a corrected cluster significance threshold of $P = 0.05$ [Forman et al., 1995; Friston, 1994; Worsley et al., 1992]. Cluster *P*-values were determined using spatial smoothness estimation in FEAT [Forman et al., 1995; Friston, 1994; Jenkinson and Smith, 2001]. The resulting statistical data, which included the combination of each participant's effect estimates and standard errors to give a mean group effect size estimate and mixed effects variance for each site, were input into a fourth-level analysis constituting the IBMA. Specifically, the intersite meta-analysis was conducted using a previously specified hierarchical model for IBMA with FLAME [Salimi-Khorshidi et al., 2009]. The site-level effect sizes and variances were modeled to provide mixed effects inference using a mixed effects group-level analysis in FLAME. The GLM included a regressor to estimate the mean effect across sites.

Aggregation of Multisite Data: Mixed Effects Model Controlling for Site

As an alternative to IBMA, we also examined a standard GLM model that controls for site. This approach could be used in contexts in which a particular effect size is not estimable at one or more sites. Group analysis was performed using FLAME (FMRIB's Local Analysis of Mixed Effects) [Behrens et al., 2003; Smith et al., 2004] with each participant's data, including parameter and variance estimates for each contrast from the lower-level analysis. The GLM for each contrast of interest included regressors for each site, age, and sex. To correct for multiple comparisons, resulting *Z*-statistic images were thresholded using clusters determined by $Z > 2.3$ and a corrected cluster significance threshold of $P = 0.05$ [Forman et al., 1995; Friston, 1994; Worsley et al., 1992]. Cluster *P*-values were determined using spatial smoothness estimation in FEAT [Forman et al., 1995; Friston, 1994; Jenkinson and Smith, 2001].

Comparison of Maps From the IBMA to Mixed Effects Model Controlling for Site

To compare the IBMA and the mixed effects model controlling for site, we used the Dice similarity measure (DSM) [Bennett and Miller, 2010; Dice, 1945], a symmetric measure of the resemblance of two binary images that has been used in previous work to measure the number of activated voxels that are shared between two fMRI images [Salimi-Khorshidi et al., 2009]. The DSM coefficient ranges from 0 (indicating no overlap) to 1 (indicating perfect overlap). *Z*-statistic activation maps from the IBMA and covariance models were first combined to create a map of the union of overlapping voxel-wise activation for the group maps. Next, a count of the number of non-zero voxels was extracted from each of the *z*-statistic maps for the IBMA, covariance model, and union map, using fslmaths. The DSM coefficient was calculated for activation during each condition of interest with the following equation:

$$\mathrm{DSM} = (2 \times |A \cap B|)/(|A| + |B|)$$

where A represents the *z*-statistic activation map from the IBMA and B represents the *z*-statistic activation map from the mixed effects model controlling for site. Thus, the DSM coefficient was calculated by dividing twice the number of overlapping voxels between the two images by the sum of the number of voxels in A and B separately.

The DSM coefficient does not distinguish between differences in spatial location versus extent; therefore, we also reported the percentage of overlapping voxels between the two approaches. For the calculation of percentage overlap, the numerator was the number of overlapping non-zero voxels between the two maps and the denominator was the number of non-zero voxels in the map with fewer non-zero voxels (because this number would represent the maximum possible number of overlapping voxels).

Testing the Effect of Site on Activation

In the aggregation study, we also conducted an ANOVA for each ROI and contrast to test whether site had a significant effect on percent signal change, controlling for age and sex. The percent signal change data were derived from the ROI analyses and were separate from the GLM approaches that controlled for site at the whole-brain level.

RESULTS

Participant Characteristics

Demographic information for the aggregation (Study 1) sample and the traveling participants (Study 2) sample is presented in Table I. For the aggregation study sample, the age of the sample differed by site ($F(7,110) = 3.83$, $P = 0.001$).

STUDY 1: TRAVELING PARTICIPANTS STUDY

Behavioral Performance in Traveling Participants Study

In the traveling participants sample, participants performed the task with a mean accuracy of 97.2% (S.D. = 4.0) across all conditions (Table II); mean RT across all conditions was 1,377 ms (S.D. = 298). Mean accuracy and RT did not differ across sites in the traveling participants study (mean accuracy: $F(7,127) = 0.65$, $P = 0.71$; mean RT: $F(7,127) = 0.40$, $P = 0.90$). In addition, there were no between-site differences for the traveling participants on accuracy or RT for any of the individual conditions (all ps

TABLE II. Mean accuracy and RT for traveling participants and aggregation study

	Overall	Emotion labeling	Emotion matching	Gender labeling	Gender matching	Match forms Shape matching
Traveling participants						
Accuracy	97.2 (4.0)	97.3 (5.2)	96.9 (5.5)	97.1 (4.8)	97.1 (6.7)	97.4 (5.7)
RT	1,377 (298)	1,511 (330)	1,817 (414)	1,303 (308)	1,193 (340)	1,056 (281)
Aggregation study						
Accuracy	96.2 (4.0)	95.9 (5.6)	93.2 (9.4)	98.0 (4.5)	96.7 (8.7)	97.3 (4.1)
RT	1,454 (258)	1,576 (276)	1,850 (372)	1,387 (278)	1,263 (328)	1,191 (249)

>0.05). There were also no differences in mean accuracy ($t(126) = -0.589$, $P = 0.56$) or mean RT ($t(126) = 0.289$, $P = 0.77$) from the first scanning day to the second scanning day. On average, scans were conducted at 12:45 pm (S.D. = 2:39; range = 8:04 am −7:40 pm). Mean scanning time did not significantly differ between subjects ($F(7,120) = 0.53$, $P = 0.81$). Mean scanning time differed between sites ($F(7,120) = 29.5$, $P < 0.0001$). Mean accuracy and RT were not associated with time of day ($P = 0.85$, $P = 0.44$, respectively). The finding that accuracy and RT did not differ between sites held when covarying for time of day.

Variance components analysis was used to determine the proportion of variance attributable to the main effects of person, site, and day; the interactions of person × site, person × day, and site × day; and the residual due to the person × site × day interaction and random error for the behavioral performance indices and for activation in each ROI. Complete variance components results for accuracy and mean RT for each condition are shown in Supporting Information Table I.

For behavioral performance, person and a person-by-site interaction accounted for a substantial portion of the variance in accuracy across conditions (Fig. 1). Moreover, the percentage of variance accounted for by person was higher than variance accounted for by site by an order of magnitude. Person accounted for the majority of variance in RT, with percentages ranging from 60 to 85%. Person-by-site variance was minimal for RT. Similar to accuracy, person accounted for substantially greater variance than site for RT. Additional plots show behavioral performance for each participant at each timepoint (see Supporting Information Fig. 2).

G-coefficients and D-coefficients were calculated to assess the relative and absolute reliability of behavioral performance, respectively (Fig. 2; Supporting Information Table I). G-coefficients ranged from $E^2_\rho = 0.69$ (shape matching) to $E^2_\rho = 0.82$ (emotion labeling) for accuracy and from $E^2_\rho = 0.95$ (shape matching) to $E^2_\rho = 0.98$ (emotion labeling, emotion matching, and gender matching) for RT. D-coefficients were highly similar. Thus, reliability was good or excellent for accuracy and RT in all conditions, indicating that behavioral performance during the emotional faces task was reliable in the current study design across scanning sites and days.

Reliability of Traveling Participants Data: Variance Components

Variance components analysis was used to determine the proportions of variance attributable to the main effects of person, site, and day; the interactions of person × site, person × day, and site × day; and the residual due to the

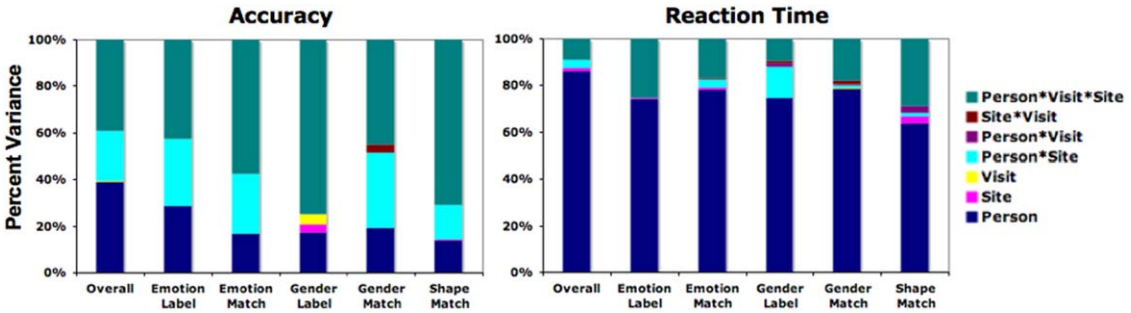

Figure 1.

Sources of variance for task performance. For both accuracy and RT, person-related factors accounted for substantially greater variance than site-related factors in all conditions of the emotional faces task.

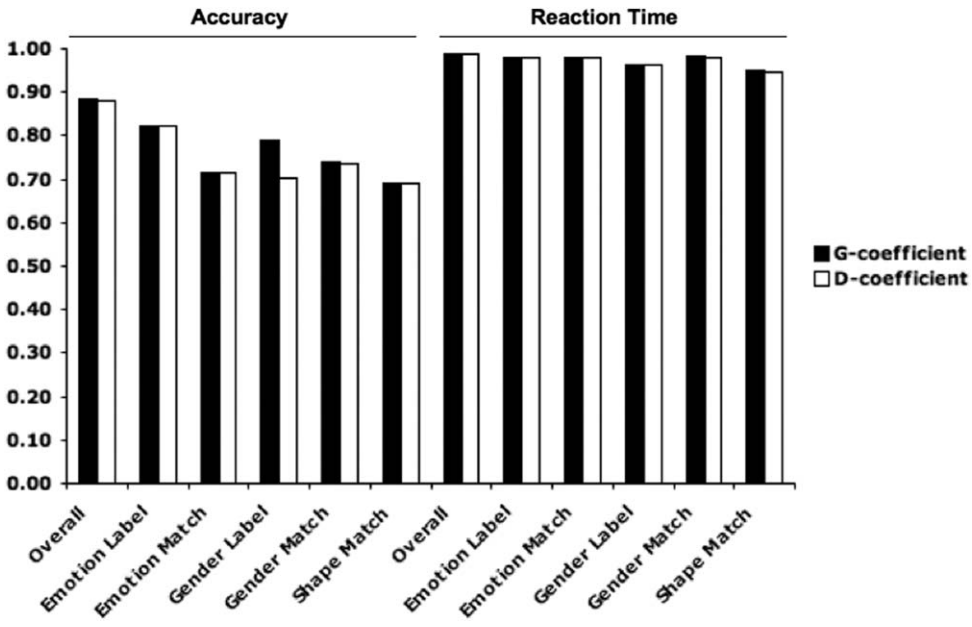

Figure 2.

Reliability of performance on the emotional faces task. G-coefficients and D-coefficients were in the good to excellent range (range: 0.69–0.98) for accuracy and RT across conditions, indicating that behavioral performance on the emotion processing task was reliable in the current study design across scanning sites and days.

person × site × day interaction and random error for mean and maximum activation in each ROI (Fig. 3). Results indicated that person-related factors accounted for a greater portion of variance compared with site-related factors across the majority of ROIs. For example, a greater proportion of variance was attributable to the person than site factor for 79% of the ROIs. On average, the proportion of variance in activation attributed to person was 11-fold larger than that attributed to site. In some cases, variance due to person was an order of magnitude higher than variance due to site. For all ROIs, the residual variance term, which includes variance due to the three-way person-by-day-by-site interaction, was the largest variance component. The person and person-by-site factors accounted for a considerable portion of the variance not attributed to the three-way interaction. Similar patterns were observed for amygdala habituation (Supporting Information Fig. 3). Overall, the proportion of variance attributed to person was larger for maximum than mean activation measures but was comparable for the anatomical and combined masks. See Supporting Information Table II for complete variance components results for each ROI and task contrast.

Variance component estimates were subsequently used to calculate G-coefficients and D-coefficients for activation in each ROI, reflecting the relative and absolute agreement of the person effect across sites and days, respectively (Fig. 4). Reliability of activation ranged from poor to excellent, depending on the ROI, task condition, and activation measure. Maximum activation measures generally showed stronger reliability than mean activation measures. For mean activation in the IFG during emotion labeling relative to implicit baseline, G-coefficients were $E_\rho^2 = 0.30$, 0.38, 0.30, and 0, for the combined right mask, anatomical right mask, combined left mask, and anatomical left mask, respectively. For maximum activation in the IFG during emotion labeling relative to implicit baseline, G coefficients were $E_\rho^2 = 0.64$, 0.45, 0.83, and 0.57 for the combined right mask, anatomical right mask, combined left mask, and anatomical left mask, respectively. Contrasts of emotion labeling compared with gender labeling or emotion matching produced similar results. D-coefficients were highly similar to G-coefficients for both mean and maximum activation in all ROIs and across contrasts (G- and D-coefficients for all regions and contrasts are shown in Supporting Information Table II).

For the amygdala, reliability of habituation was generally higher than mean (but not maximum) activation. Reliability was typically higher for the right than left amygdala for both activation and habituation. For habituation (mean) in the right amygdala, G-coefficients were $E_\rho^2 = 0.71$ for the combined mask and $E_\rho^2 = 0.25$ for the

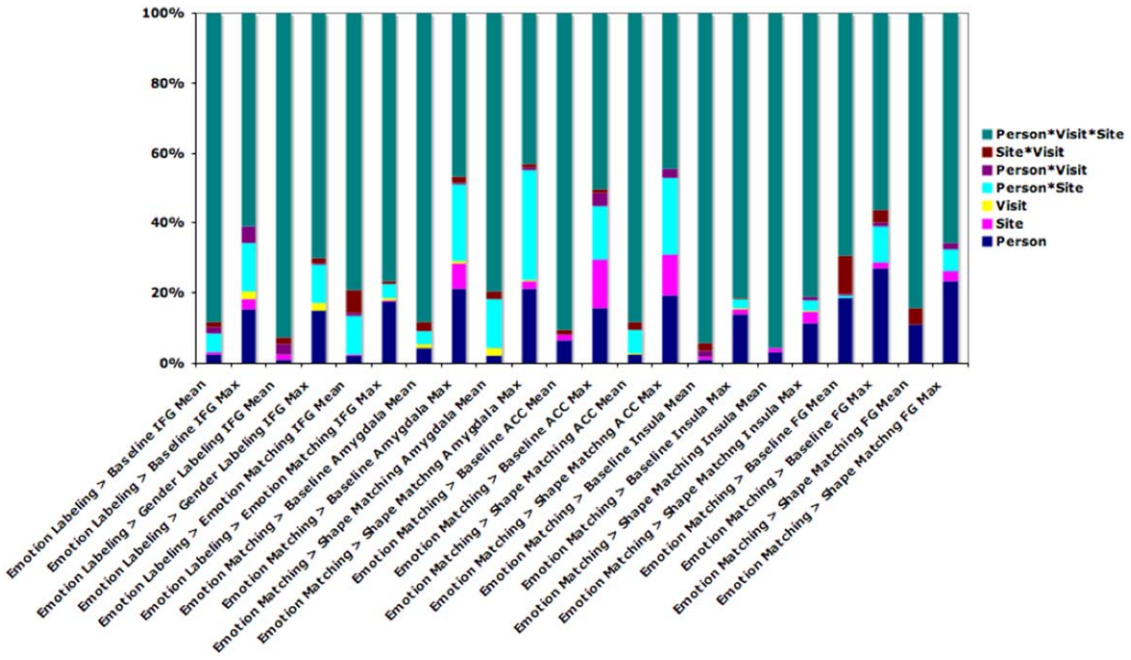

Figure 3.

Sources of variance for functional activation. Variance components analysis of activation measures revealed that variance due to person-related factors was substantially greater than variance due to site. A person-by-day-by-site interaction accounted for considerable variance in activation across ROI. Person-related factors accounted for greater variance in maximum than mean activation measures.

anatomical mask. Similar results were observed for habituation during emotion matching relative to shape matching. For mean activation in the right amygdala, G-coefficients were $E_\rho^2 = 0.41$ for the combined mask and $E_\rho^2 = 0.44$ for the anatomical mask during emotion matching relative to implicit baseline. For maximum activation in the right amygdala, G-coefficients were $E_\rho^2 = 0.74$ for the combined mask and $E_\rho^2 = 0.82$ for the anatomical mask during emotion matching relative to implicit baseline. Although reliability was high for maximum measures of amygdala activation across all contrasts, G-coefficients were more variable for mean activation depending on the contrast. Reliability was generally higher for emotion matching when compared with implicit baseline than when compared with shape matching.

During emotion matching relative to implicit baseline, G-coefficients for the ACC were $E_\rho^2 = 0.55$ and 0.52 for mean activation and $E_\rho^2 = 0.76$ and 0.61 for maximum activation in the left and right anatomical masks, respectively. G-coefficients for the insula were $E_\rho^2 = 0.42$, 0, 0, and 0 for mean activation and $E_\rho^2 = 0.47$, 0.49, 0.79, and 0.85 for maximum activation in the combined left, combined right, anatomical left, and anatomical right masks, respectively. G-coefficients for the fusiform gyrus were $E_\rho^2 = 0.75$, 0.83, 0.79, and 0.82 for mean activation and $E_\rho^2 = 0.62$, 0.91, 0.78, and 0.88 for maximum activation in the combined left, combined right, anatomical left, and anatomical right masks, respectively. Results were similar for the contrast of emotion matching versus shape matching.

Results indicate that the relative ranking of persons was reliable, or generalizable, in the current study design across scanning sites and days for core emotion processing ROIs during relevant task conditions. Specifically, generalizability ICCs reflected the reliability of activation for the amygdala during emotion matching (typically higher for habituation than amplitude of mean activation) and the IFG during emotion labeling. Strong generalizability was also evident for the ACC, insula, and fusiform gyrus during emotion matching. Consistent with the greater proportion of variance attributed to person for maximum activation measures, the relative ranking of persons was more generalizable (with the majority in the good or excellent range) for maximum activation measures than mean activation measures across ROIs.

A GLM tested whether the effects of site, day, or a site-by-day interaction were significant for the amygdala and IFG for each contrast. For mean activation, the effect of site was significant for 1 of 20 measures, day for 1 of 20, and site × day for 1 of 20. For maximum activation, the effect of site was significant for 3 of 20 measures, day for 4

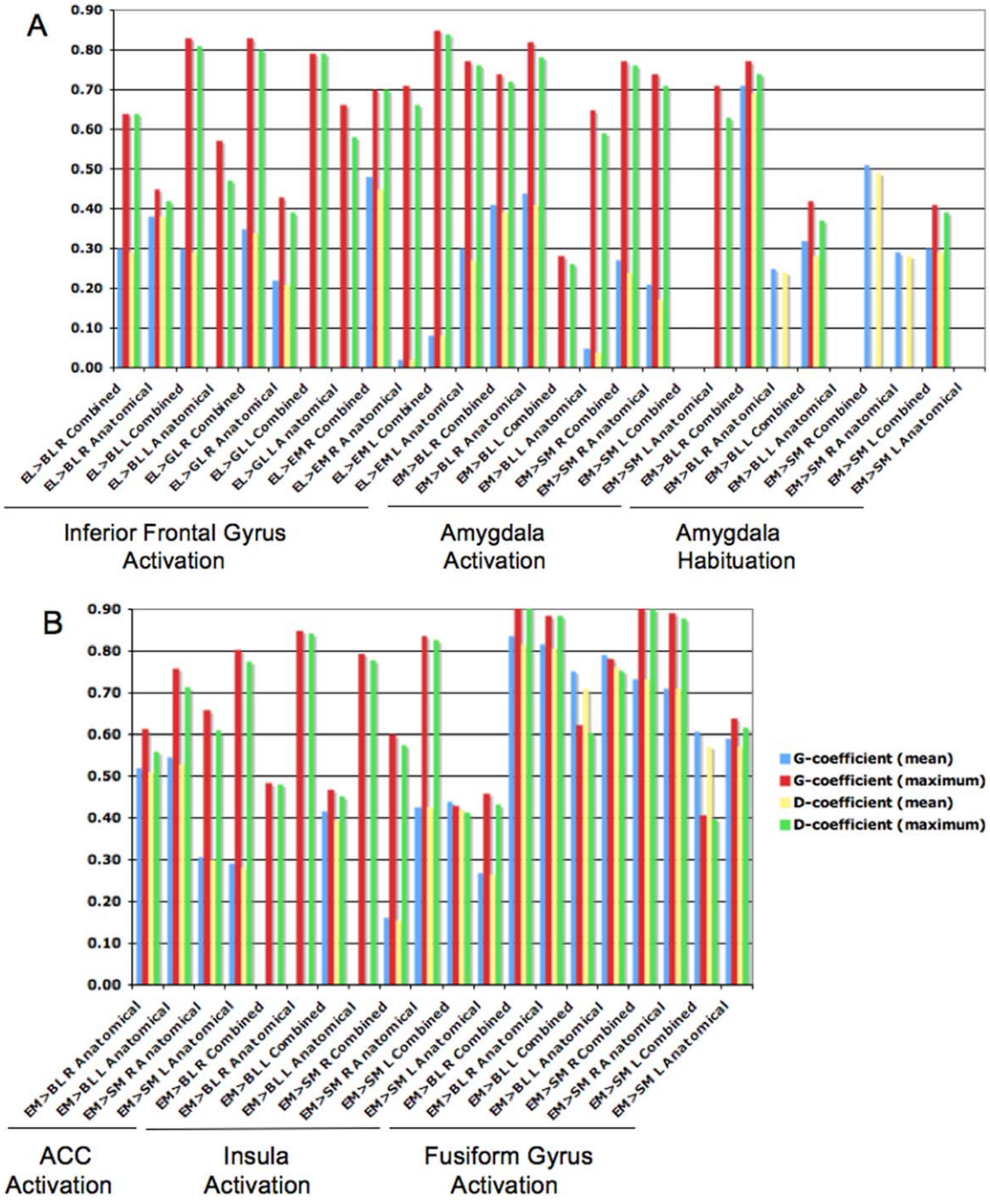

Figure 4.

Reliability for functional activation. G-coefficients and D-coefficients showed that the reliability of activation varied by region and condition. **A**) During emotion labeling versus implicit baseline, G-coefficients for the IFG ranged from $E_\rho^2 = 0$ to $E_\rho^2 = 0.38$ for mean activation and from $E_\rho^2 = 0.45$ to $E_\rho^2 = 0.83$ for maximum activation. During emotion matching versus implicit baseline, G-coefficients for the right amygdala ranged from $E_\rho^2 = 0$ to $E_\rho^2 = 0.44$ for mean activation and from $E_\rho^2 = 0.28$ to $E_\rho^2 = 0.82$ for maximum activation, and $E_\rho^2 = 0.25$ to $E_\rho^2 = 0.71$ for habituation (mean) and $E_\rho^2 = 0$ to $E_\rho^2 = 0.77$ for habituation (maximum). **B**) During emotion matching versus implicit baseline, G-coefficients ranged from $E_\rho^2 = 0.52$ to $E_\rho^2 = 0.55$ for mean activation and from $E_\rho^2 = 0.61$ to $E_\rho^2 = 0.76$ for maximum activation for the ACC, from $E_\rho^2 = 0$ to $E_\rho^2 = 0.42$ for mean activation and from $E_\rho^2 = 0.47$ to $E_\rho^2 = 0.85$ for maximum activation for the insula, and from $E_\rho^2 = 0.75$ to $E_\rho^2 = 0.83$ for mean activation and from $E_\rho^2 = 0.62$ to $E_\rho^2 = 0.91$ for maximum activation for the fusiform gyrus.

of 20, and site × day for 0 of 20 (Supporting Information Table III for specific regions and contrasts). These results suggest that activation measures rarely differed significantly between sites or between the first and second scan days.

STUDY 2: AGGREGATION STUDY

Behavioral Performance in the Aggregation Study

Participants performed the task with a mean accuracy of 96.2% (S.D. = 4.0) across all conditions in the sample used in the study of data aggregation (Table II). Mean RT across

TABLE III. Regions of activation for IBMA and mixed effects model controlling for site in the aggregation study

	Image-based meta-analysis				Mixed effects covariance model			
Activation	Region	Z max	Peak voxel	Cluster size	Region	Z max	Peak voxel	Cluster size
Emotion Labeling > Baseline	Lateral occipital cortex, occipital pole, amygdala, thalamus (bilateral)	23	−24, −94, −2	21,937	Lateral occipital cortex, occipital pole (bilateral)	14	30, −92, −2	17,109
	Superior parietal lobe (left)	8.58	−28, −50, 42	6,195	Precentral gyrus, middle frontal gyrus, inferior frontal gyrus, orbitofrontal cortex (left)	5.98	−44, 6, 24	3,531
	Precentral gyrus, middle frontal gyrus, inferior frontal gyrus, orbitofrontal cortex (right)	9.2	44, 8, 34	3,739	Precentral gyrus, middle frontal gyrus, inferior frontal gyrus, orbitofrontal cortex (right)	7.45	48, 8, 44	3,169
	Paracingulate gyrus, superior frontal gyrus (bilateral)	7.51	−6, 12, 48	812	Amygdala, thalamus (left)	7.42	−24, −28, −2	2,019
	Frontal pole (bilateral)	6.82	6, 52, 28	398	Amygdala, thalamus (right)	7.29	24, −28, −2	1,457
	Superior parietal lobule (right)	6.26	26, −56, 44	359	Paracingulate gyrus, superior frontal gyrus (bilateral)	5.49	−6, 12, 44	515
					Superior parietal lobule (left)	6.59	−28, −56, 46	457
Emotion Matching > Baseline	Lateral occipital cortex, occipital pole, amygdala, thalamus (bilateral)	24.3	−26, −94, 0	53,441	Lateral occipital cortex, occipital pole, amygdala, thalamus (bilateral)	13	26, −90, −2	50,573
	Paracingulate gyrus, supplementary motor cortex, middle frontal gyrus, inferior frontal gyrus (bilateral)	9.41	−8, 10, 46	1,340	Paracingulate gyrus, supplementary motor cortex, middle frontal gyrus, inferior frontal gyrus (bilateral)	8.77	−2, 8, 52	1,163
	Frontal pole (right)	6.53	10, 64, 26	390				
Emotion Labeling > Emotion Matching	Postcentral gyrus, central opercular cortex, Heschl's gyrus, insula (left)	5.56	−56, −14, 18	2,108	Postcentral gyrus, central opercular cortex, Heschl's gyrus, insula (left)	5.19	−58, −28, 10	2,922

TABLE III. (continued).

	Image-based meta-analysis				Mixed effects covariance model			
Activation	Region	Z max	Peak voxel	Cluster size	Region	Z max	Peak voxel	Cluster size
	Postcentral gyrus, central opercular cortex, Heschl's gyrus, insula (right)	4.91	46, −8, 6	1,174	Postcentral gyrus, central opercular cortex, Heschl's gyrus, insula (right)	4.48	52, −16, 16	1,624
	Orbitofrontal cortex, inferior frontal gyrus (left)	6.94	−60, 28, −6	755	Medial frontal cortex, paracingulate gyrus (bilateral)	4.47	−2, 48, −10	434
	Medial frontal cortex, paracingulate gyrus (bilateral)	6.45	0, 48, −10	425	Orbitofrontal cortex, inferior frontal gyrus (left)	3.96	−52, 20, −6	373
	Precentral gyrus	4.89	2, −24, 50	385				
	Lingual gyrus	3.94	−30, −52, 4	188				
Emotion Labeling > Gender Labeling	Inferior frontal gyrus, orbitofrontal cortex (right)	7.15	52, 24, 8	2,844	Inferior frontal gyrus, orbitofrontal cortex (right)	7.38	52, 24, 8	2,210
	Inferior frontal gyrus, orbitofrontal cortex (left)	6.02	−54, 14, 14	1,394	Middle temporal gyrus, lateral occipital cortex (right)	5.87	46, −54, 6	940
	Middle temporal gyrus, lateral occipital cortex (right)	5.74	50, −44, 6	1,131	Inferior frontal gyrus, orbitofrontal cortex (left)	4.65	−54, 12, 10	666
	Middle temporal gyrus, lateral occipital cortex (left)	5.3	−56, −62, 6	543	Supramarginal gyrus, middle temporal gyrus, lateral occipital cortex (left)	4.58	−68, −44, 20	478
	Superior frontal gyrus, paracingulate gyrus (bilateral)	4.43	4, 20, 52	388				
Emotion Matching > Shape Matching	Lateral occipital cortex, occipital pole, middle temporal gyrus, angular gyrus, fusiform gyrus, thalamus, amygdala (bilateral)	22.2	−26, −94, 0	49,307	Lateral occipital cortex, occipital pole, middle temporal gyrus, angular gyrus, fusiform gyrus, thalamus, amygdala (bilateral)	13.3	−6, −98, −4	28,499
	Paracingulate gyrus, supplementary motor cortex, middle frontal gyrus, superior frontal gyrus, inferior frontal gyrus, orbitofrontal cortex, precentral gyrus (bilateral)	8.95	−6, 20, 44	2,264	Inferior frontal gyrus, middle frontal gyrus, precentral gyrus, orbitofrontal cortex (left)	8.04	−52, 18, 28	6,251
					Inferior frontal gyrus, middle frontal gyrus, precentral gyrus, orbitofrontal cortex (right)	8.94	50, 32, 12	5,453

TABLE III. (continued).

	Image-based meta-analysis				Mixed effects covariance model			
Activation	Region	Z max	Peak voxel	Cluster size	Region	Z max	Peak voxel	Cluster size
					Paracingulate gyrus, superior frontal gyrus, supplementary motor cortex (bilateral)	6.77	−4, 18, 48	1,592
					Superior parietal lobule, lateral occipital cortex (left)	6.15	30, −60, 44	1,532
					Superior parietal lobule, lateral occipital cortex (right)	6.47	−34, −54, 44	1,321

all conditions was 1,454 ms (S.D. = 258). Mean accuracy and RT did not differ across sites (mean accuracy: $F(7,110) = 0.53$, $P = 0.81$; mean RT: $F(7,110) = 0.67$, $P = 0.70$). Nonparametric statistical tests (Kruskal-Wallis) confirmed that mean accuracy and RT did not differ between sites ($P = 0.95$, $P = 0.76$, respectively). In addition, there were no between-site differences on accuracy or RT for any of the individual conditions (all ps > 0.05).

Aggregation of Multisite Data

Results of the hierarchical model for IBMA and the mixed effects model controlling for site both showed robust activation in expected regions during the emotional faces task. Moreover, there was high correspondence between the approaches in the spatial localization of activation in cluster-based results (Fig. 5; Table III for complete details of regions). For both methods, emotion labeling relative to implicit baseline was associated with increased activation in IFG, orbitofrontal gyrus, lateral occipital cortex, occipital pole, amygdala, thalamus, superior parietal lobe, precentral gyrus, middle frontal gyrus, paracingulate gyrus, and superior frontal gyrus. During emotion matching relative to implicit baseline, activation in the amygdala, thalamus, paracingulate gyrus, supplementary motor cortex, middle frontal gyrus, IFG, occipital pole, and lateral occipital cortex was increased using both methods. For emotion labeling and emotion matching relative to implicit baseline, increased activation in frontal pole was uniquely observed in the IBMA. Directly contrasting emotion labeling with emotion matching resulted in activation in regions such as the IFG, orbitofrontal cortex, postcentral gyrus, central opercular cortex, and insula across both approaches. The IBMA analysis also showed activation in the lingual gyrus and precentral gyrus.

Higher-order contrasts of the emotion conditions with control conditions showed more circumscribed clusters of activation than the comparisons with implicit baseline. Results were highly similar for the IBMA model and the mixed effects model controlling for site. Specifically, emotion labeling versus gender labeling was associated with activation in the IFG, orbitofrontal cortex, middle temporal gyrus, and lateral occipital cortex. The IBMA analysis also showed activation in the superior frontal gyrus and paracingulate gyrus. For emotion matching versus shape matching, both approaches showed activation in the amygdala, thalamus, fusiform gyrus, angular gyrus, middle temporal gyrus, lateral occipital cortex, IFG, orbitofrontal cortex, and paracingulate gyrus. The mixed effects model covarying for site also showed activation in the superior parietal lobule.

The DSM coefficient was used to quantify voxelwise overlap between the thresholded images produced by the two approaches for each contrast. Consistent with the high correspondence between approaches in the spatial localization of activation in cluster-based results, DSM coefficients for each contrast revealed a high degree of overlap between the non-zero voxels in the maps from each approach (Table IV). For example, the DSM coefficient for the comparison of the IBMA and mixed effects model was 0.84 for emotion labeling and 0.91 for emotion matching (relative to implicit baseline). Given that the DSM coefficient can range from 0 to 1.0, with 1.0 indicating perfect similarity between two sets, this demonstrates a high degree of similarity in results from the two methods of multisite data aggregation. In the aggregation study, we tested whether site significantly related to mean percent signal change for any of the ROIs for each contrast using

TABLE IV. DSM coefficients for aggregation across site

Contrast	DSM
Emotion labeling > Baseline	0.84
Emotion matching > Baseline	0.91
Emotion labeling > Emotion matching	0.55
Emotion labeling > Gender labeling	0.81
Emotion matching > Shape matching	0.87

an ANOVA that controlled for age and sex. The site effect was significant for 2 of the 20 mean activation measures and 2 of the 20 maximum activation measures (Supporting Information Table IV for specific regions and contrasts).

DISCUSSION

Given notable contributions of multisite neuroimaging and the limited extant research about the reliability of multisite fMRI in the domain of emotion processing, this study aimed to characterize the reliability of BOLD activation during a commonly used emotional faces fMRI task and compare two statistical methods for aggregating data across sites. Variance components analysis from the traveling participants study showed that person-related variation in BOLD signal was substantially greater than site-related variation for both emotion labeling and emotion matching. On average, the proportion of variance in activation attributed to person was 11-fold larger than that attributed to site. Though the reliability of activation varied from poor to excellent depending on the region and task contrast, generalizability and dependability ICCs demonstrated the reliability of the person effect for the amygdala, IFG, ACC, insula, and fusiform gyrus during the primary conditions in which they were expected to be recruited. In the aggregation study, fMRI data for all healthy individuals recruited as control participants in the NAPLS study were aggregated across sites using IBMA and a mixed effects model controlling for site. Robust activation of regions such as the amygdala and IFG was observed in both approaches, and the DSM coefficient confirmed a high degree of spatial overlap in results. Taken together, the present work suggests that it is possible to obtain reliable signal and robust activation in core emotion processing regions during relevant task conditions in the present multisite design and supports the use of the emotional faces task in future neuroimaging studies that would benefit from the advantages of a multisite investigation.

Using reliability analyses within a G-theory framework for the traveling participants sample, the present results suggested that activation during emotion processing was reliable across scanning sites and testing days. Consistent with prior studies examining variance components of the BOLD response [Brown et al., 2011; Costafreda, 2009; Gountouna et al., 2010; Yendiki et al., 2010], variance components analysis demonstrated that person-related variability was greater than site-related variability across the majority of ROIs. In some cases, variance due to the person effect was an order of magnitude higher than variance due to the site effect. These results suggest that person-related variability can be large enough and site-related variability small enough for the effective aggregation of data across sites. G-coefficients and D-coefficients demonstrated substantial variation in the reliability of the person effect for activation depending on the ROI, task condition, and measure of activation. In this study, person-related variance was substantially higher for maximum percent signal change than for mean percent signal change. A higher ICC and the ability to generalize beyond a given site and testing day depend on sufficient variability in the measure of interest. Consistently, maximum measures of activation may ensure more variation because the voxel of maximum change can differ between scans and participants, whereas mean activation is averaged across the same voxels at each measurement. Several prior studies of multisite reliability have observed comparable reliability for maximum percent signal change in other cognitive domains [Costafreda et al., 2007; Friedman et al., 2008; Yendiki et al., 2010]. Although fMRI studies have traditionally relied on mean percent signal change, the current findings suggest that maximum percent signal change may be a more reliable measure for the emotional faces task. In addition, masks defined anatomically and masks defined by the intersection of anatomical structures with functional maps yielded similar results for variance components estimates and the reliability of the person effect for activation.

Reliability is frequently higher in regions that are more robustly activated by an fMRI task [Bennett and Miller, 2010; Brown et al., 2011; Caceres et al., 2009]. Thus, we focused on the amygdala during emotion matching and the IFG during emotion labeling for this study, consistent with prior research showing robust activation during these task conditions in healthy controls [Hariri et al., 2000; Lieberman et al., 2007]. To inform research using other emotion tasks, we also examined the ACC, insula, and fusiform gyrus during emotion matching. The present task robustly activated the fusiform gyrus and insula, but not the ACC. Nevertheless, our findings suggest that BOLD signal in these regions was consistent across scanning sites and testing days. Though prior research on the reliability of BOLD signal during emotion-related tasks in multisite studies is limited, an investigation using a sad affect paradigm showed a stronger contribution of the person effect to variance in activation for the prefrontal cortex than the amygdala [Suckling et al., 2008]. Cognitive tasks often show lower signal reliability relative to motor and sensory tasks [Bennett and Miller, 2010]; thus, the nature of emotion processing tasks might limit reliability. Nevertheless, the current results are encouraging in that they demonstrate that it is possible to achieve fair to excellent reliability in various regions central to emotion processing during specific task conditions.

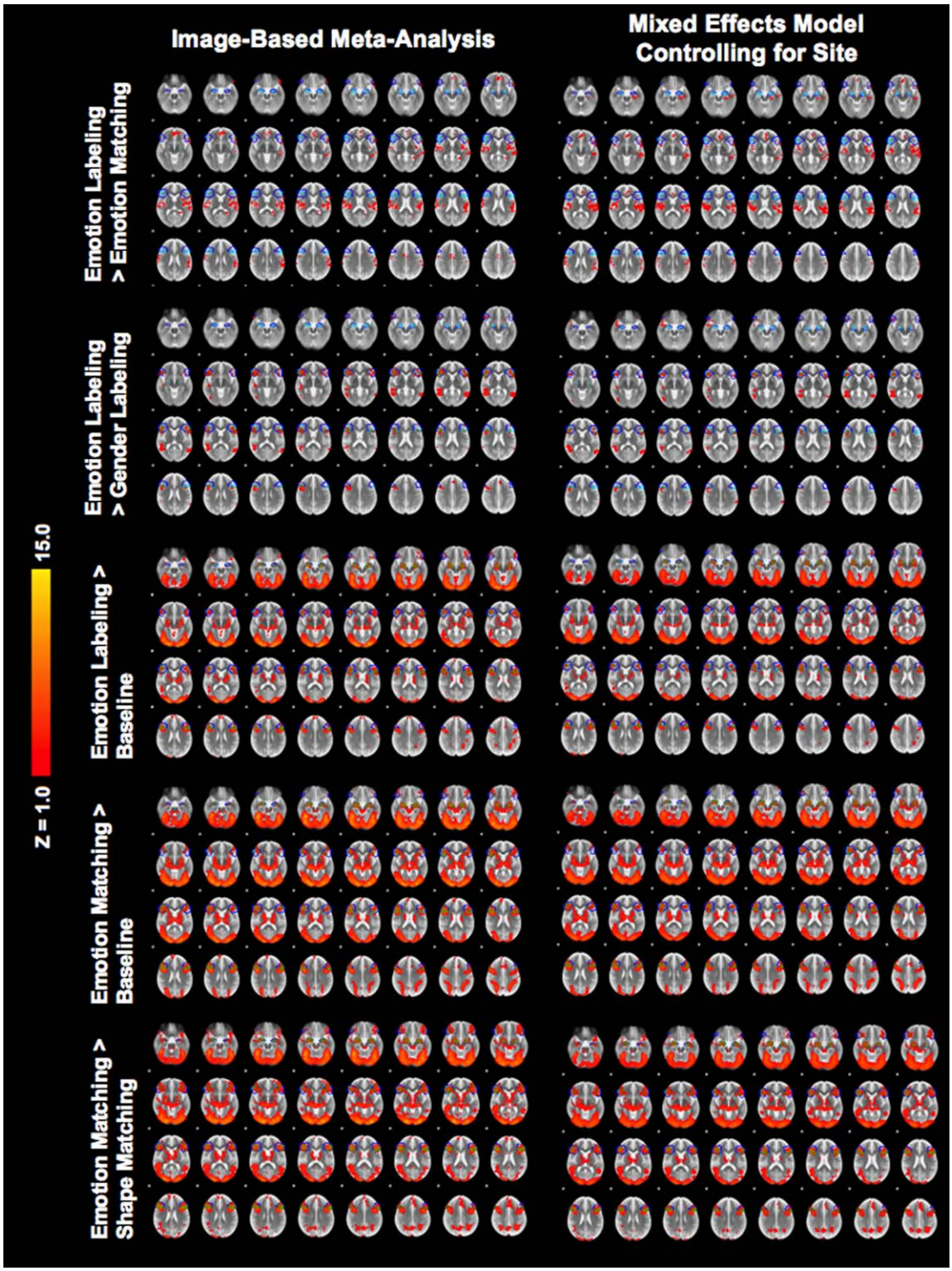

Figure 5.

Functional activation maps in the aggregation study. The IBMA and mixed effects model controlling for site produced similar results for functional activation during emotion processing. Effects of activation were robust in regions previously identified in single-site studies using this task, namely the amygdala and IFG. Outlines of the ROIs for the amygdala and IFG are overlaid (anatomical masks in dark blue, combined functional and anatomical masks in light blue).

Prior single-site studies of the test-retest reliability of the amygdala have found limited reliability for activation, though ICCs values have ranged from poor to excellent depending on the study, contrast, and population. Based on these prior single-site studies [Johnstone et al., 2005; Lipp et al., 2014; Plichta et al., 2012; Sauder et al., 2013; Van den Bulk et al., 2013], the reliability coefficients for amygdala activation in this study fall within the expected range. Activation was more stable in the right amygdala than the left amygdala, consistent with prior work using the same task (Manuck et al., 2007; Plichta et al., 2014), which may be due to high visuospatial demands of emotion matching rather than underlying neural properties. It has also been suggested that amygdala activation is more reliable to fearful faces than other facial expressions (Sauder et al., 2013; Stark et al., 2004); however, the current block design precluded comparisons by emotional expression. Importantly, the present multisite investigation found that amygdala habituation was a more reliable measure than mean activation, consistent with a prior single-site study using the same task [Plichta et al., 2014]. These findings point to the importance of examining habituation in the amygdala in future studies using emotional face paradigms.

Another fundamental aspect of multisite evaluation is the extent to which multisite results replicate functional activation in regions that are spatially consistent and known to be functionally important for the task based on single-site studies. In the second study, both the IBMA and mixed effects model controlling for site produced robust activation consistent with regions expected based on prior studies. Specifically, emotion labeling was found to increase activation in the IFG, consistent with prior evidence that this region plays a critical role in labeling emotions and dampening amygdala activation during the task [Lieberman et al., 2007]. Moreover, emotion matching increased activation in the amygdala, consistent with numerous studies demonstrating that the emotion matching condition robustly activates the amygdala [Fakra et al., 2008; Gee et al., 2012; Hariri et al., 2000; Klumpp et al., 2012; Lieberman et al., 2007]. These findings suggest that the emotional faces task produces consistent results for activation when data are aggregated across different sites with different scanners. Comparing methods for aggregating data is also an important step to ensure that statistical approaches are valid. Prior research comparing image-based versus coordinate-based methods for aggregating fMRI data across sites demonstrated a clear advantage for the IBMA method for minimizing information loss while accounting for differences between sites; however, it may not be possible to conduct a hierarchical analysis of group maps generated for every site (e.g., if one site contributes control participants but few or no patient participants). Using a mixed effects model controlling for site would allow data from such a site to be aggregated without requiring an underpowered case-control contrast. Results from the IBMA and mixed effects model controlling for site in this study demonstrated a high degree of overlap (DSM = 0.84 for emotion labeling; DSM = 0.91 for emotion matching, relative to implicit baseline), and both exhibited robust effects of BOLD response in expected task-related regions. The present findings suggest that controlling for site within a mixed effects model may provide an alternative approach that yields similar results for use in specific cases that render an IBMA impractical. Thus, the present findings suggest that either model would provide a valid method for aggregating data in the NAPLS study and may inform future methods for the aggregation of fMRI data across multiple scanning sites.

The present findings may also facilitate strategic decisions regarding the design of future multisite fMRI investigations. In addition to substantial variance due to participant effects, the findings from the traveling participants study indicated a high percentage of variance due to an interaction between person and site, as well as to a three-way interaction between person, site, and day. The overall pattern of the present results, with substantial contributions of participant and interactions between participant, site, and day, with little contribution due to site, replicates prior findings on the reliability of activation during working memory tasks [Brown et al., 2011; Yendiki et al., 2010]. Similarly, prior studies of fMRI reliability have reported that error constituted the majority of variance [e.g., Brown et al., 2011; Suckling et al., 2008; Yendiki et al., 2010]. The present results and use of G-theory provide greater specificity to examine the interactions underlying the greatest portion of variance. The substantial contribution of a participant-by-site interaction and a participant-by-site-by-day interaction to the variance in the present investigation suggests that the rank ordering of activation among individual participants differed across sites and scanning sessions within a single site. However, the greater magnitude of variance due to the participant-by-site interaction, relative to site alone, indicates that variability between data acquired at different sites was due more to individual participants producing different BOLD responses on different occasions rather than to overall site differences. Substantial variability in working memory paradigms has also been attributed to a participant-by-site interaction, with numerous possible explanations [Brown et al., 2011; Forsyth et al., 2014; Yendiki et al., 2010]. For example, it may be that procedures (e.g., participant placement) or magnet stability changed over time at different sites, or that participants exhibited individual differences in the consistency of their performance across sites. Although it may be that participants responded differently across time due to learning effects, this possibility is limited in this study by the fact that site order was counterbalanced across participants. Although the effect of day was generally nonsignificant, it might also be that a greater test-retest interval between scans would reduce the possibility of habituation effects. Future studies may reduce participant-by-site variance through rigorous efforts to maintain the same procedures and ensure

compliance with the task. Moreover, because the effects of site and day were rarely significant, it is likely that the variance attributed to the participant-by-site-by-day interaction relates to factors that differed by site and day within participants, which are often difficult to measure (e.g., variation in attention and arousal, caffeine, sleep, diet, nicotine, time of day of the scan). A future study that controls for these factors would provide insight into this variance, which is likely to be person-related. Although high unexplained variance is a concern of fMRI studies generally, results from the current and prior studies suggest that results from standard tasks are often similar across scanners and can nevertheless yield important insights into differences in brain activation between control and patient groups.

Behavioral results demonstrated that participants performed with a high level of accuracy overall and for each condition, consistent with prior findings from the same task [Lieberman et al., 2007]. Moreover, results showed consistent performance for accuracy and RT across sites. Variance components analysis of the behavioral measures demonstrated high person-related variance and extremely low site-related variance, with variance due to participant especially high for RT. Moreover, generalizability and dependability ICCs demonstrated good to excellent reliability of the person effect for both accuracy and RT across conditions. The data do not suggest that there were considerable effects of practice or learning despite the repetition of the task, which may suggest that the task is well-suited for longitudinal or within-participant designs.

Several limitations of this study warrant further consideration in future work. Two runs of fMRI data were collected during each scan in this study; however, prior work suggests that averaging across additional runs increases reliability [Brown et al., 2011; Friedman et al., 2008]. Given the limited number of trials for each condition, reliability in this study may have been higher with more fMRI data. In addition, measures of spatial overlap such as the DSM will be limited by the threshold used in the imaging analyses. As in previous investigations of multisite fMRI [e.g., Brown et al., 2011; Costafreda, 2009; Gountouna et al., 2010; Suckling et al., 2008; Sutton et al., 2008], the ROIs used in this study were created based on predefined anatomical structures or group-level functional results. The possibility that ROI masks did not contain all locations of activation in some cases, or contained too many voxels in other cases, may have limited the ability to capture within-participant, across-session variability and thus contributed to the substantial effect of the participant-by-site-by-day interaction. Though it would be ideal to use participant-specific anatomy and functional maps to define ROIs, the present approach allowed us to examine the reliability of fMRI in the context of standard practice. We conducted secondary analyses to test the accuracy of the backwarping of standard anatomical ROIs onto each participant's individual anatomy and observed a high degree of alignment. Despite advantages of G-theory methods for assessing reliability in fMRI, it is important to note that this approach assumes that site is a random effect and aims to measure how well activation would generalize from the case when all subjects are scanned at a different single scanner to the case when all subjects are scanned on all possible scanners. Finally, the feasibility of having a larger sample of participants travel to each of the NAPLS sites was limited. Thus, the sample size of the traveling participant component of the study was relatively small, and demographic variance between the traveling participants may have been low relative to the general population. Given the essential role of variance in determining reliability, oversampling of similar individuals can lead to underestimates of reliability coefficients. It may be that confidence intervals on reliability measures included zero in the current investigation. Future studies with more participants will be useful to further assess between-site variability. It is also important to note that multisite investigations do not necessarily provide higher power per se because the power may be reduced due to the addition of site-related variance [e.g., Suckling et al., 2008]. Thus, multisite studies might require oversampling participants to account for the added variance of site. In this study, power analyses showed that, on average, the multisite analysis required 2.8% more participants to detect an effect than at a single site (See Supporting Information Results). This finding is consistent with prior research suggesting that approximately 10% more participants are needed for a multisite study [Suckling et al., 2008].

CONCLUSION

The present multisite investigation of an emotional faces task demonstrates the feasibility of aggregating fMRI data acquired across multiple sites to produce robust examinations of activation. Multiple methods for combining data across sites produced encouraging results of multisite effects that are consistent with prior single-site findings, and these results may inform future approaches to analyzing multisite fMRI data. Moreover, measures of activation exhibited substantially greater person-related variance than site-related variance, suggesting that person-related variability is adequate to compensate for site-related differences. In addition, these findings extend the investigation of multisite reliability to emotion processing, which is likely to be a critical area of study as brain imaging continues to expand in the domains of social cognition and clinical neuroscience. Despite variation in the reliability of BOLD signal depending on the region and condition, when taken together, the present results show that it is possible to observe robust and reliable activation for core emotion processing regions during relevant task conditions in a multisite investigation. In conjunction with the current task's widespread use with both healthy controls [Creswell et al., 2007; Lieberman et al., 2007; Payer et al., 2012] and clinical populations to date [Fakra et al., 2008;

Foland-Ross et al., 2012; Gee et al., 2012; Kircanski et al., 2012; Payer et al., 2011], the present findings suggest that the multisite implementation of the emotional faces task would be appropriate and valuable to the field. The present findings have important implications for the future design of multisite neuroimaging studies, which are likely to facilitate scientific discovery in the study of large samples that would be difficult to recruit at a single site, such as unique clinical populations.

REFERENCES

Addington J, Cadenhead KS, Cannon TD, Cornblatt B, McGlashan TH, Perkins DO, Seidman LJ, Tsuang M, Walker EF, Woods SW, Heinssen R (2007): North American prodrome longitudinal study: A collaborative multisite approach to prodromal schizophrenia research. Schizophr Bull 33:665–672.

Barch DM, Mathalon DH (2011): Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: Psychometric and quality assurance considerations. Biol Psychiatry 70:13–18.

Beckett LA, Harvey DJ, Gamst A, Donohue M, Kornak J, Zhang H, Kuo JH, Alzheimer's Disease Neuroimaging Initiative (2010): The Alzheimer's disease neuroimaging initiative: Annual change in biomarkers and clinical outcomes. Alzheimers Dement 6:257–264.

Behrens TEJ, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CAM, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, Thompson AJ, Brady JM, Matthews PM (2003): Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nat Neurosci 6:750–757.

Bennett CM, Miller MB (2010): How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 1191:133–155.

Bernal-Casas D, Balaguer-Ballester E, Gerchen MF, Iglesias S, Walter H, Heinz A, Meyer-Lindenberg A, Stephan KE, Kirsch P (2013): Multi-site reproducibility of prefrontal-hippocampal connectivity estimates by stochastic DCM. Neuroimage 82C: 555–563.

Blackford JU, Allen AH, Cowan RL, Avery SN (2013): Amygdala and hippocampus fail to habituate to faces in individuals with an inhibited temperament. Soc Cogn Affect Neurosci 8:143–150.

Brandt DJ, Sommer J, Krach S, Bedenbender J, Kircher T, Paulus FM, Jansen A (2013): Test-retest reliability of fMRI brain activity during memory encoding. Front Psychiatry 4:163.

Brennan RL (2001): Generalizability Theory. Springer, New York.

Brown GG, Mathalon DH, Stern H, Ford J, Mueller B, Greve DN, McCarthy G, Voyvodic J, Glover G, Diaz M, Yetter E, Ozyurt IB, Jorgensen KW, Wible CG, Turner JA, Thompson WK, Potkin SG (2011): Multisite reliability of cognitive BOLD data. Neuroimage 54:2163–2175.

Caceres A, Hall DL, Zelaya FO, Williams SCR, Mehta MA (2009): Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45:758–768.

Cannon TD, Cadenhead K, Cornblatt B, Woods SW, Addington J, Walker E, Seidman LJ, Perkins D, Tsuang M, McGlashan T, Heinssen R (2008): Prediction of psychosis in youth at high clinical risk: A multisite longitudinal study in North America. Arch Gen Psychiatry 65:28–37.

Casey BJ, Cohen JD, O'Craven K, Davidson RJ, Irwin W, Nelson CA, Noll DC, Hu X, Lowe MJ, Rosen BR, Truwitt CL, Turski PA (1998): Reproducibility of fMRI results across four institutions using a spatial working memory task. Neuroimage 8:249–261.

Cicchetti DV, Sparrow SA (1981): Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. Am J Ment Defic 86:127–137.

Costafreda SG (2009): Pooling FMRI data: Meta-analysis, mega-analysis and multi-center studies. Front Neuroinform 3:33.

Costafreda SG, Brammer MJ, Vêncio RZN, Mourão ML, Portela LAP, De Castro CC, Giampietro VP, Amaro E Jr (2007): Multisite fMRI reproducibility of a motor task using identical MR systems. J Magn Reson Imaging 26:1122–1126.

Creswell JD, Way BM, Eisenberger NI, Lieberman MD (2007): Neural correlates of dispositional mindfulness during affect labeling. Psychosom Med 69:560–565.

Di Nocera F, Ferlazzo F, Borghi V (2001): G Theory and the reliability of psychophysiological measures: A tutorial. Psychophysiology 38:796–806.

Dice LR (1945): Measures of the amount of ecologic association between species. Ecology 26:297–302.

Fakra E, Salgado-Pineda P, Delaveau P, Hariri AR, Blin O (2008): Neural bases of different cognitive strategies for facial affect processing in schizophrenia. Schizophr Res 100:191–205.

First MB, Spitzer RL, Gibbon M, Wiliams JBW (2002): Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P) New York: Biometrics Research, New York State Psychiatric Institute.

Foland-Ross LC, Bookheimer SY, Lieberman MD, Sugar CA, Townsend JD, Fischer J, Torrisi S, Penfold C, Madsen SK, Thompson PM, Altshuler LL (2012): Normal amygdala activation but deficient ventrolateral prefrontal activation in adults with bipolar disorder during euthymia. Neuroimage 59:738–744.

Ford JM, Roach BJ, Jorgensen KW, Turner JA, Brown GG, Notestine R, Bischoff-Grethe A, Greve D, Wible C, Lauriello J, Belger A, Mueller BA, Calhoun V, Preda A, Keator D, O'Leary DS, Lim KO, Glover G, Potkin SG, Mathalon DH (2009): Tuning in to the voices: A multisite FMRI study of auditory hallucinations. Schizophr Bull 35:58–66.

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

Forsyth JK, McEwen SC, Gee DG, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos HW, Tsuang MT, van Erp TGM, Walker EF, Hamann S, Woods SW, Qiu M, Cannon TD (2014): Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: Analysis from the North American prodrome longitudinal study. Neuroimage 97:41–52.

Friedman L, Glover GH (2006): Report on a multicenter fMRI quality assurance protocol. J Magn Reson Imaging 23:827–839.

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

Friston KJ (1994): Functional and effective connectivity in neuroimaging: A synthesis. Hum Brain Mapp 2:56–78.

Fusar-Poli P, Placentino A, Carletti F, Landi P, Allen P, Surguladze S, Benedetti F, Abbamonte M, Gasparotti R, Barale F, Perez J, McGuire P, Politi P (2009): Functional atlas of emotional faces processing: A voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J Psychiatry Neurosci 34:418–432.

Gee DG, Karlsgodt KH, Van Erp TGM, Bearden CE, Lieberman MD, Belger A, Perkins DO, Olvet DM, Cornblatt BA, Constable T, Woods SW, Addington J, Cadenhead KS, McGlashan TH, Seidman LJ, Tsuang MT, Walker EF, Cannon TD (2012): Altered age-related trajectories of amygdala-prefrontal circuitry in adolescents at clinical high risk for psychosis: A preliminary study. Schizophr Res 134:1–9.

Glover GH, Mueller BA, Turner JA, Van Erp TGM, Liu TT, Greve DN, Voyvodic JT, Rasmussen J, Brown GG, Keator DB, Calhoun VD, Lee HJ, Ford JM, Mathalon DH, Diaz M, O'Leary DS, Gadde S, Preda A, Lim KO, Wible CG, Stern HS, Belger A, McCarthy G, Ozyurt B, Potkin SG (2012): Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J Magn Reson Imaging 36: 39–54.

Gountouna V-E, Job DE, McIntosh AM, Moorhead TWJ, Lymer GKL, Whalley HC, Hall J, Waiter GD, Brennan D, McGonigle DJ, Ahearn TS, Cavanagh J, Condon B, Hadley DM, Marshall I, Murray AD, Steele JD, Wardlaw JM, Lawrie SM (2010): Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49:552–560.

Gradin V, Gountouna V-E, Waiter G, Ahearn TS, Brennan D, Condon B, Marshall I, McGonigle DJ, Murray AD, Whalley H, Cavanagh J, Hadley D, Lymer K, McIntosh A, Moorhead TW, Job D, Wardlaw J, Lawrie SM, Steele JD (2010): Between- and within-scanner variability in the CaliBrain study n-back cognitive task. Psychiatry Res 184:86–95.

Hariri AR, Bookheimer SY, Mazziotta JC (2000): Modulating emotional responses: Effects of a neocortical network on the limbic system. Neuroreport 11:43–48.

Huang L, Wang X, Baliki MN, Wang L, Apkarian AV, Parrish TB (2012): Reproducibility of structural, resting-state BOLD and DTI data between identical scanners. PLoS ONE 7:e47684.

Jenkinson M, Smith S (2001): A global optimisation method for robust affine registration of brain images. Med Image Anal 5: 143–156.

Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Johnstone T, Somerville LH, Alexander AL, Oakes TR, Davidson RJ, Kalin NH, Whalen PJ (2005): Stability of amygdala BOLD response to fearful faces over multiple scan sessions. Neuroimage 25:1112–1123.

Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, Williamson D, Ryan N (1997): Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 36:980–988.

Kennedy DN, Lange N, Makris N, Bates J, Meyer J, Caviness VS Jr (1998): Gyri of the human neocortex: An MRI-based analysis of volume and variance. Cereb Cortex 8:372–384.

Kim DI, Mathalon DH, Ford JM, Mannell M, Turner JA, Brown GG, Belger A, Gollub R, Lauriello J, Wible C, O'Leary D, Lim K, Toga A, Potkin SG, Birn F, Calhoun VD (2009): Auditory oddball deficits in schizophrenia: An independent component analysis of the fMRI multisite function BIRN study. Schizophr Bull 35:67–81.

Kircanski K, Lieberman MD, Craske MG (2012): Feelings into words: Contributions of language to exposure therapy. Psychol Sci 23:1086–1091.

Klumpp H, Angstadt M, Phan KL (2012): Shifting the focus of attention modulates amygdala and anterior cingulate cortex reactivity to emotional faces. Neurosci Lett 514:210–213.

Kober H, Feldman Barrett L, Joseph J, Bliss-Moreau E, Lindquist K, Wager T (2008): Functional grouping and cortical-subcortical interactions in emotion: A meta-analysis of neuroimaging studies. NeuroImage 42:998–1031.

Kring AM, Moran EK (2008): Emotional response deficits in schizophrenia: Insights from affective science. Schizophr Bull 34: 819–834.

Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RSE, Davis SM, Davis CE, Lebowitz BD, Severe J, Hsiao JK (2005): Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New Engl J Med 353: 1209–1223.

Lieberman MD, Eisenberger NI, Crockett MJ, Tom SM, Pfeifer JH, Way BM (2007): Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. Psychol Sci 18:421–428.

Lieberman MD, Inagaki TK, Tabibnia G, Crockett MJ (2011): Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. Emotion 11:468–480.

Lipp I, Murphy K, Wise RG, Caseras X (2014): Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses. Neuroimage 86:335–342.

Makris N, Meyer JW, Bates JF, Yeterian EH, Kennedy DN, Caviness VS (1999): MRI-Based topographic parcellation of human cerebral white matter and nuclei II. Rationale and applications with systematics of cerebral connectivity. Neuroimage 9:18–45.

Manuck SB, Brown SM, Forbes EE, Hariri AR (2007): Temporal stability of individual differences in amygdala reactivity. Am J Psychiatry 164:1613–1614.

McGlashan TH, Miller TJ, Woods SW, Hoffman RE, Davidson L (2001): Instrument for the assessment of prodromal symptoms and states. In: Miller T, Mednick SA, McGlashan TH, Libiger J, Johannessen JO, editors. Early Intervention in Psychotic Disorders. NATO Science Series 91. Springer Netherlands. pp 135–149. Available at: http://link.springer.com/chapter/10.1007/978-94-010-0892-1_7.

Meyer-Lindenberg A (2012): The future of fMRI and genetics research. NeuroImage 62:1286–1292.

Mueser KT, Doonan R, Penn DL, Blanchard JJ, Bellack AS, Nishith P, DeLeon J (1996): Emotion recognition and social competence in chronic schizophrenia. J Abnorm Psychol 105: 271–275.

Ojemann JG, Buckner RL, Akbudak E, Snyder AZ, Ollinger JM, McKinstry RC, Rosen BR, Petersen SE, Raichle ME, Conturo TE (1998): Functional MRI studies of word-stem completion: Reliability across laboratories and comparison to blood flow imaging with PET. Hum Brain Mapp 6:203–215.

Payer DE, Lieberman MD, London ED (2011): Neural correlates of affect processing and aggression in methamphetamine dependence. Arch Gen Psychiatry 68:271–282.

Payer DE, Baicy K, Lieberman MD, London ED (2012): Overlapping neural substrates between intentional and incidental down-regulation of negative emotions. Emotion 12:229–235.

Pearlson G (2009): Multisite collaborations and large databases in psychiatric neuroimaging: Advantages, problems, and challenges. Schizophr Bull 35:1–2.

Phan KL, Wager T, Taylor SF, Liberzon I (2002): Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. neuroimage 16:331–348.

Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes ABM, Sauer C, Tost H, Esslinger C, Colman P, Wilson F, Kirsch P, Meyer-Lindenberg A (2012): Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. Neuroimage 60:1746–1758.

Plichta MM, Grimm O, Morgen K, Mier D, Sauer C, Haddad L, Tost H, Esslinger C, Kirsch P, Schwarz AJ, Meyer-Lindenberg A (2014): Amygdala habituation: A reliable fMRI phenotype. Neuroimage 103C:383–390.

Potkin SG, Turner JA, Brown GG, McCarthy G, Greve DN, Glover GH, Manoach DS, Belger A, Diaz M, Wible CG, Ford JM, Mathalon DH, Gollub R, Lauriello J, O'Leary D, Van Erp TGM, Toga AW, Preda A, Lim KO (2009): Working memory and DLPFC inefficiency in schizophrenia: The FBIRN study. Schizophr Bull 35:19–31.

Salimi-Khorshidi G, Smith SM, Keltner JR, Wager TD, Nichols TE (2009): Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. Neuroimage 45:810–823.

Sauder CL, Hajcak G, Angstadt M, Phan KL (2013): Test-retest reliability of amygdala response to emotional faces. Psychophysiology 50:1147–1156.

Shavelson RJ, Webb NM (1991): Generalizability Theory: A Primer. SAGE, Thousand Oaks, CA.

Shrout PE, Fleiss JL (1979): Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004): Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23:S208–S219.

Stark R, Schienle A, Walter B, Kirsch P, Blecker C, Ott U, Schafer A, Sammer G, Zimmerman M, Vaitl D (2004): Hemodynamic effects of negative emotional pictures—A test-retest analysis. Neuropsychobiology 50:108–118.

Suckling J, Ohlssen D, Andrew C, Johnson G, Williams SCR, Graves M, Chen C-H, Spiegelhalter D, Bullmore E (2008): Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp 29:1111–1122.

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

Tabibnia G, Lieberman MD, Craske MG (2008): The lasting effect of words on feelings: Words may facilitate exposure effects to threatening images. Emotion 8:307–317.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C (2009): The NimStim set of facial expressions: Judgments from untrained research participants. Psychiatry Res 168:242–249.

Van den Bulk BG, Koolschijn PCMP, Meens PHF, Van Lang NDJ, Van der Wee NJA, Rombouts SARB, Vermeiren RRJM, Crone EA (2013): How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Dev Cogn Neurosci 4:65–76.

Van Horn JD, Toga AW (2009): Multisite neuroimaging trials. Curr Opin Neurol 22:370–378.

Voyvodic JT (2006): Activation mapping as a percentage of local excitation: fMRI stability within scans, between scans and across field strengths. Magn Reson Imaging 24:1249–1261.

Webb NM, Shavelson RJ (2005): Generalizability theory: Overview. In: Encyclopedia of Statistics in Behavioral Science. Wiley, Chichester. Available at: http://onlinelibrary.wiley.-com/doi/10.1002/0470013192.bsa703/abstract.

Wechsler D (1999): Wechsler Abbreviated Scale of Intelligence. New York, NY: The Psychological Corporation: Harcourt Brace & Company.

Woolrich MW, Ripley BD, Brady M, Smith SM (2001): Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370–1386.

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

Wurnig MC, Rath J, Klinger N, Höllinger I, Geissler A, Fischmeister FP, Aichhorn M, Foki T, Kronbichler M, Nickel J, Siedentopf C, Staffen W, Verius M, Golaszewski S, Koppelstätter F, Knosp E, Auff E, Felber S, Seitz RJ, Beisteiner R (2013): Variability of clinical functional MR imaging results: A multicenter study. Radiology 268:521–531.

Yendiki A, Greve DN, Wallace S, Vangel M, Bockholt J, Mueller BA, Magnotta V, Andreasen N, Manoach DS, Gollub RL (2010): Multi-site characterization of an fMRI working memory paradigm: Reliability of activation indices. Neuroimage 53:119–131.

You X, Adjouadi M, Guillen MR, Ayala M, Barreto A, Rishe N, Sullivan J, Dlugos D, Vanmeter J, Morris D, Donner E, Bjornson B, Smith ML, Bernal B, Berl M, Gaillard WD (2011): Sub-patterns of language network reorganization in pediatric localization related epilepsy: A multisite study. Hum Brain Mapp 32:784–799.

Zou KH, Greve DN, Wang M, Pieper SD, Warfield SK, White NS, Manandhar S, Brown GG, Vangel MG, Kikinis R, Wells WM III (2005): Reproducibility of functional MR imaging: Preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 237: 781–789.