NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 25 (2005) 1112–1123

Stability of amygdala BOLD response to fearful faces over multiple scan sessions

Tom Johnstone,[a,b,*] Leah H. Somerville,[a] Andrew L. Alexander,[a] Terrence R. Oakes,[a] Richard J. Davidson,[a,b,c] Ned H. Kalin,[a,b,c] and Paul J. Whalen[a,b,c]

[a] *W.M. Keck Laboratory for Functional Brain Imaging and Behavior, University of Wisconsin, WI 53705, USA*
[b] *Department of Psychiatry, University of Wisconsin, WI 53705, USA*
[c] *Department of Psychology, University of Wisconsin, WI 53705, USA*

Received 28 July 2004; revised 1 December 2004; accepted 6 December 2004
Available online 5 March 2005

We used fMRI to examine amygdala activation in response to fearful facial expressions, measured over multiple scanning sessions. 15 human subjects underwent three scanning sessions, at 0, 2 and 8 weeks. During each session, functional brain images centered about the amygdala were acquired continuously while participants were shown alternating blocks of fearful, neutral and happy facial expressions. Intraclass correlation coefficients calculated across the sessions indicated stability of response in left amygdala to fearful faces (as a change from baseline), but considerably less left amygdala stability in responses to neutral expressions and for fear versus neutral contrasts. The results demonstrate that the measurement of fMRI BOLD responses in amygdala to fearful facial expressions might be usefully employed as an index of amygdala reactivity over extended periods. While signal change to fearful facial expressions appears robust, the experimental design employed here has yielded variable responsivity within baseline or comparison conditions. Future studies might manipulate the experimental design to either amplify or attenuate this variability, according to the goals of the research.

Keywords: Stability; Amygdala; BOLD response; Fearful facial expression

Introduction

A number of fMRI studies have demonstrated increased activation of the amygdala to the presentation of biologically relevant stimuli, in particular fearful facial expressions (Breiter et al., 1996; Irwin et al., 1996; Kim et al., 2003; Morris et al., 1996; Phillips et al., 1998; Whalen et al., 1998a,b, 2001). Little, however, is known about the stability of amygdala BOLD responses to fearful facial expressions during multiple scan sessions over extended periods of time (e.g., weeks or months). An understanding of amygdala response stability is crucial in longitudinal studies such as those relating amygdala activation to long-term changes of mood in normal subjects, clinical trials of treatments for a variety of psychopathological disorders (cf. Schwartz and Rauch, 2004) or genetic or other biological factors (e.g., Hariri and Weinberger, 2003). At the least, lack of test–retest reliability due to random variation in amygdala activation over time would limit the sensitivity to time-dependent changes of interest. A potentially greater problem would be systematic changes in amygdala activation over time, which would complicate comparisons between different experimental groups (e.g., treatment versus control) across time.

The reproducibility of fMRI results depends upon a number of subject-dependent variables. For example, the psychological state of the subject can vary across scan sessions, in both unpredictable and predictable ways. Of particular relevance to studies of amygdala activation to emotional facial expressions is the subject's anxiety at the time of the scan, which has been shown to correlate positively with BOLD response to neutral faces (Somerville et al., 2004). Further variability in BOLD contrast between scan sessions is likely to result from learning related to the experimental task and stimuli (e.g., habituation). Some studies have measured habituation effects in fMRI, including studies of amygdala response to emotional stimuli (Fischer et al., 2003; Wright et al., 2001). Most of these studies have examined within-session effects, rather than effects over multiple scan sessions.

There have been relatively few brain imaging studies that have reported test–retest reliability, and most of those studies have addressed reliability across scans within a single scan session. For example, Tegeler et al. (1999) calculated the reliability of BOLD activation across three scan runs of a finger-opposition task measured on a 4 T MRI scanner. Other studies have measured reliability over longer time frames, but are limited to simple motor or visual stimulation tasks, or analyses of data from a single subject. An example of the latter is an fMRI study of BOLD response in

* Corresponding author. W.M. Keck Laboratory for Functional Brain Imaging and Behavior, Waisman Center, 1500 Highland Avenue, Madison, WI 53705, USA. Fax: +1 608 265 8737.
E-mail address: itjohnstone@wisc.edu (T. Johnstone).

Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.12.016

motor, visual and cognitive tasks measured in a single subject over 33 scan sessions (McGonigle et al., 2000), and a related follow-up study on the same data set (Smith et al., in press). The authors found that intersession variability was not large compared to within-session variability, although it is unclear how this result would generalize to multiple subjects. One study that examined fMRI reliability in multiple subjects over multiple sessions was a study of activation during a working memory task (Manoach et al., 2001). Subjects were scanned twice with a mean inter-scan interval of approximately 14 weeks, on a 1.5 T scanner. Percent signal change in the voxel with the maximum t statistic within three areas involved in working memory was used to calculate intraclass correlation coefficients (ICC; Shrout and Fleiss, 1979) as indices of test–retest reliability. Used in the context of test–retest reliability, the ICC is a measure of the ratio of between-subjects variance to total variance, which includes both between-subjects variance and between-tests variance. The ICC will thus approach 1 when the variability across subjects is much larger than the variability within-subjects across repeated measurements. The ICCs ranged from 0.81 in dorsolateral prefrontal cortex, to 0.68 and 0.49 in intraparietal sulcus and insula, respectively. Thus, moderate to high reliability was found in these brain regions. The study was also directly relevant to the current study because percent signal change values from a priori hypothesized regions of interest (ROI) were used to test reliability, using intraclass correlation coefficients. The results can thus be seen as independent of the somewhat arbitrary setting of significance levels used in whole-brain voxelwise statistical comparisons, which can be misleading in studies of reproducibility (Smith et al., in press).

To our knowledge, there has been only one previous study of test–retest reliability of brain imaging data from the amygdala over extended periods of time. Schaefer et al. (2000) measured the test–retest reliability over 6 months of PET measures of resting regional metabolic rate of glucose (rCMR) in a number of subcortical structures, including the amygdala, hippocampus, thalamus and the anterior caudate nucleus. They found reliability in left amygdala but not right amygdala, a result that they tentatively suggested might be due to the effects of variability in anxiety across different scan sessions on right amygdala metabolism. The extent to which these data bear on measures of amygdala BOLD activation in functional MRI experiments is difficult to gauge.

Imaging the amygdala with fMRI presents particular difficulties due to signal dropout caused by intravoxel dephasing, which is a function of large differences in magnetic susceptibility between brain matter and proximal sinuses. Signal dropout will lead to a generally lower signal to noise ratio (SNR), which will reduce the reproducibility of BOLD responses in the amygdala. A further problem is that slight differences in the position of the head in the scanner from one scan session to another will change the amount of signal dropout at specific loci within the amygdala, as well as the average SNR across the amygdala. To maximize SNR, it is thus imperative to use a scan sequence that mimimizes the deleterious effects of magnetic susceptibility inhomogeneity. In our laboratory, we have adopted a coronal oblique, partial brain acquisition centered about the amygdala that affords the best whole amygdala coverage on our GE 3 T scanner, relatively free of susceptibility artifacts and dropout (see e.g., Kim et al., 2003, 2004; Somerville et al., 2004). A similar acquisition has been independently confirmed as optimal for imaging amygdala (Chen et al., 2003). Here, we present average SNR images for the amygdala region to facilitate the comparison of reproducibility data in future studies.

Analysis of fMRI data involves a number of preprocessing steps, most notably motion correction and temporal filtering, that reduce noise and thus will increase signal reliability, although their efficacy will depend upon the specific implementation used, the relative merits of which are beyond the scope of this article (but see Gold et al., 1998). While spatial filtering should also lead to increased reliability (due in part to it mitigating the effects of small residual differences in brain position between successive scans), the amount of spatial blurring applied (indeed, whether or not any spatial blurring is used at all) depends on the expected volume of activation in a given experiment. In particular, for studies concerning the amygdala in which only small regions might be activated, over-smoothing of the images will likely lead to a decrease in sensitivity and reliability due to partial volume effects. In addition, spatial smoothing will tend to obscure potentially interesting, small-scale differences in the localization of functional activations. A similar argument applies to studies that use an ROI approach; the size and shape of the extracted ROI will affect the reliability. Given these concerns about the appropriate use of spatial smoothing, and the appropriate selection of ROIs, the effects of both these factors on fMRI sensitivity and reliability were examined in this study.

We used fMRI to study amygdala BOLD activation in response to the presentation of fearful facial expressions, measured over three scanning sessions at 0, 2 and 8 weeks. We also assessed the reliability of amygdala response to neutral faces, which have typically been used as a comparison condition for fear expressions (but which vary in ways that might make them less suitable for longitudinal studies; see Somerville et al., 2004). Individual-subject data were analyzed using a general linear model and estimated contrast values from amygdala ROIs for all subjects were then analyzed for test–retest reliability. The effects of spatial smoothing and ROI selection on fMRI sensitivity and reliability were examined.

Materials and methods

Participants

15 human subjects (age range 21–51, mean age 33 years, 13 female) underwent three scanning sessions, at 0, 2 and 8 weeks. All subjects provided informed written consent before participation. This group of subjects served as the control group in a longitudinal study of treatment for Generalized Anxiety Disorder. All were screened for DSM-IV axis I and II diagnosis and had Hamilton Anxiety (HAM-A) scores below 5. All subjects provided informed written consent. This investigation was conducted in accordance with the guidelines of the Human Subjects Committee of the University of Wisconsin-Madison.

Procedure

One week prior to the first scan session, subjects attended a 30-min fMRI simulation session within a mock scanner. Any concerns or questions about the experimental procedure were answered, and subjects underwent a simulated scan in which they lay in the mock scanner while examples of the types of images to be used in the experiment (although not the actual images to be used) were shown. In addition, simulated noise of the scanner was presented through headphones. The simulation session was designed to familiarize the subjects with the MRI procedure, and to reduce any initial apprehension or anxiety. During this session, a dental mold was made to be used as a custom bite bar during the scanning sessions.

To assess subject anxiety immediately prior to the experimental session, all subjects completed the Hamilton Rating Scale for Anxiety (Ham-A), Penn State Worry Questionnaire (PSWQ) and State-trait Anxiety Inventory (STAI). Subjects were then carefully placed in the scanner and asked to make themselves as comfortable as possible while gently biting on the bite bar. The Avotech goggle system used to present visual stimuli was then adjusted to provide a clear view of a test image with both eyes. Padding was arranged around the subject's head, which together with use of the bite bar served to minimize head movement and ensure as much as possible that the subject's head was positioned the same way across scan sessions.

During each of the three scan sessions, participants were shown alternating 18 s blocks of fearful (F), neutral (N) and happy (H) facial expressions during two scan runs, with the relative order of happy and fear blocks counterbalanced within and between subjects (we present only the results for the fearful and neutral stimuli here). Each scan run started and finished with an 18 s fixation (+) baseline block, thus a typical scan would be as follows: +, N, H, N, F, N, H, N, F, N+.

Each block consisted of six repetitions of six identities (3 female) from a standardized stimulus set (Ekman and Friesen, 1976; identities used were PE, SW, WF, PF, C, GS). The same stimuli were used for all 3 scan sessions. All stimuli were standardized for contrast and luminance. Each expression was displayed for 200 ms, with an inter-trial interval (ITI) of 300 ms consisting of a fixation cross on a black background (i.e., 2 faces/s).

Image acquisition

A 3 T SIGNA (General Electric Medical Systems) MRI scanner with a quadrature head coil and high speed gradients was used to acquire both whole brain, axial, high resolution anatomical scans (3D SPGR; 240 mm FOV, 256 × 192 in-plane resolution; 124 slices, 1.1 mm slice thickness) and functional gradient echo EPI scans. 18 partial brain (amygdala centered) coronal oblique functional slices were obtained (3 mm slice thickness; 0.5 mm interslice gap; 64 × 64 in-plane resolution; 180 mm FOV; 108 3D volumes per scan run; TR/TE/Flip = 2000 ms/30 ms/60°). This slice acquisition has been used extensively in our laboratory and previous studies by the authors because it minimizes through-plane dephasing, phase cancellation and phase dispersion, and thus results in data relatively free of susceptibility artifacts and dropout. A similar acquisition scheme has recently been proposed as optimal for amygdala imaging by Chen et al. (2003). Fig. 1 shows the orientation and position of acquired slices.

Image analysis

All data processing was performed using AFNI software (Cox, 1996), with the exception of coregistration of data from different sessions and normalization to Talairach space, for which FLIRT software (Jenkinson et al., 2002) was used. Individual subject data were motion corrected, low pass filtered (cutoff = 0.15 Hz), and were then analyzed using a general linear model (GLM) with separate regressors for each expression type, formed by convolving a stimulus boxcar function with an ideal hemodynamic response function (HRF). The GLM yielded a set of contrast maps (fearful versus baseline, neutral versus baseline, fearful versus neutral) for each individual.

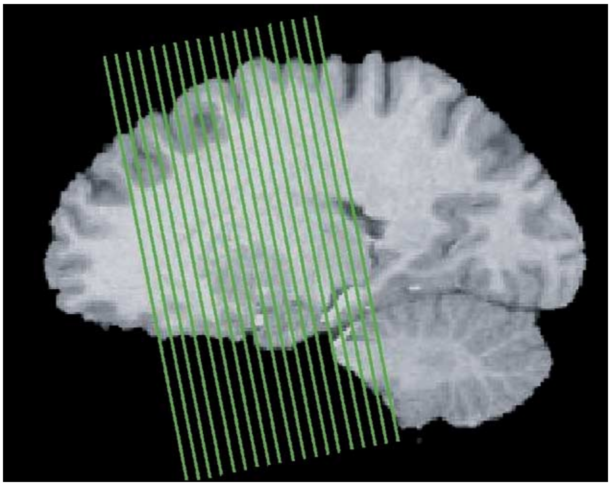

Fig. 1. Sagittal section through left amygdala showing position and orientation of 18 acquired coronal oblique slices.

Two different voxelwise indices of activation were then calculated for each subject and each contrast; percent signal change estimates (derived by dividing each voxel's contrast estimate by the estimated baseline for that voxel and multiplying by 100) and contrast z scores. Although percent signal change is the usual manner in which to quantify image signal changes in functional MRI experiments, this metric can give spuriously high values in regions of large signal dropout (such as amygdala or ventromedial prefrontal cortex). A number of studies have used contrast z scores when quantifying signal change in such brain regions. Z scores are essentially a ratio of contrast to noise, and thus have a slightly different interpretation than percent signal change. We decided to test the extent to which the use of z scores in the amygdala resulted in different sensitivity or reliability than percent signal change.

Activation maps were normalized into Talairach Space (Talairach and Tournoux, 1988) using FLIRT. To examine the effect of spatially blurring the data on test–retest reliability, we then applied a Gaussian spatial blur with a full width at half maximum (FWHM) of 4 mm. The unfiltered data sets were also analyzed.

The activation indices were extracted from the same Talairach-defined amygdala regions of interest (ROIs) for all subjects, and formed the basis for subsequent analyses. In addition, right and left amygdala ROIs were defined on the basis of t tests applied to contrast maps from the first scan session. Activation indices were then extracted from these ROIs for all three scan sessions. The use of such a statistically defined ROI from the first session allows an assessment of how well results from the first scan session could be reproduced in subsequent sessions.

To test the statistical significance of main contrast effects (i.e., those that remain stable over scan sessions), as well as session effects (i.e., changes in contrasts across scan sessions), estimated contrast indices from Talairach-defined and statistically defined amygdala ROIs for all subjects were entered into a mixed effects analysis, with subjects as a random factor and scan session as a fixed factor.

To quantify test–retest reliability, intraclass correlations were calculated for extracted contrast indices from both Talairach-defined

Table 1
Mean reported anxiety across the scan sessions

	Ham-A	PSWQ	STAI-state
Scan session 1	1.3 (1.0)	31.1 (5.4)	43.7 (4.9)
Scan session 2	1.3 (1.2)	31.0 (3.7)	43.9 (4.6)
Scan session 3	1.8 (1.6)	31.9 (4.3)	44.4 (4.9)

Numbers in parentheses are standard deviations.

and statistically defined amygdala ROIs for all subjects across all three scan sessions, as well as across pairs of scan sessions. Two types of ICCs were calculated. Single measure ICC is a measure of the repeatability of a single measurement, in this case an estimate of response to fear faces in a single scan session. Such an indicator of test–retest reliability is relevant to studies that measure a quantity at only one time, or that use the results measured at one time to predict future results or constrain future analyses. Average measure ICC is an indication of the reliability of the mean of repeated measures (i.e., the means of estimates of fear response over two or three scan sessions), and would be relevant to longitudinal studies where the reliability of mean responses in a control or placebo group over multiple scan sessions will be determined.

Results

Anxiety ratings data

Means and standard deviations of rated anxiety on the Ham-A, PSWQ and STAI-state scales are presented in Table 1. Repeated measures ANOVA showed no significant difference across scan session for any of the scales (each measure $F(2,26) < 1$), indicating that there was no consistent group-level difference in reported anxiety across scan sessions. In subsequent analyses of amygdala activation across the three scan sessions, we calculated the correlation between the anxiety measures and measures of activation (as well as their relative changes over scan sessions), and found no significant correlations. Note that individuals showed a limited range in anxiety change across scan sessions, with the largest individual range in Ham-A of 4, in STAI-state of 8 and in PSWQ of 8. The range of anxiety scores across subjects was also small, presumably due to the pre-screening criteria for inclusion in the study. Thus, given the restricted range of anxiety scores in this pre-screened sample of subjects, the lack of correlation between measures of anxiety and amygdala response to facial expressions is not surprising.

SNR measurements

As can be seen in Fig. 2, all voxels lying within the Talairach-defined amygdala ROIs had a mean SNR of greater than 50, with lower values in the ventral and medial portions. The mean SNR for the left and right amygdalae was 77.2 (SD = 0.38) and 83.2 (SD = 0.27), respectively, although as can be seen in Fig. 3, the SNR values were skewed towards higher values in the right, compared to the left, amygdala. These SNR values compare favorably to recent studies of optimized scan parameters for EPI imaging of the amygdala (Chen et al., 2003; Robinson et al., 2004). SNR values can most intuitively be thought of in terms of percent signal change. An SNR of 50 equates to an RMS noise level of about 2% baseline signal, which makes it difficult to measure small signal differences, although this limitation can be overcome by choosing an appropriate experimental design and by measuring a large enough group of participants. Note that the SNR we report here is that of data that have not been spatially blurred. Spatial blurring will have the effect of increasing the effective SNR, at the potential cost of reduced ability to identify and localize very small, focal activations.

It is worth emphasizing that although a large number of studies have been able to detect differences in amygdala response under different experimental conditions, the low SNR attainable in this part of the brain excludes the interpretation of null results. That is, when studies fail to find activation in the amygdala, the cause is

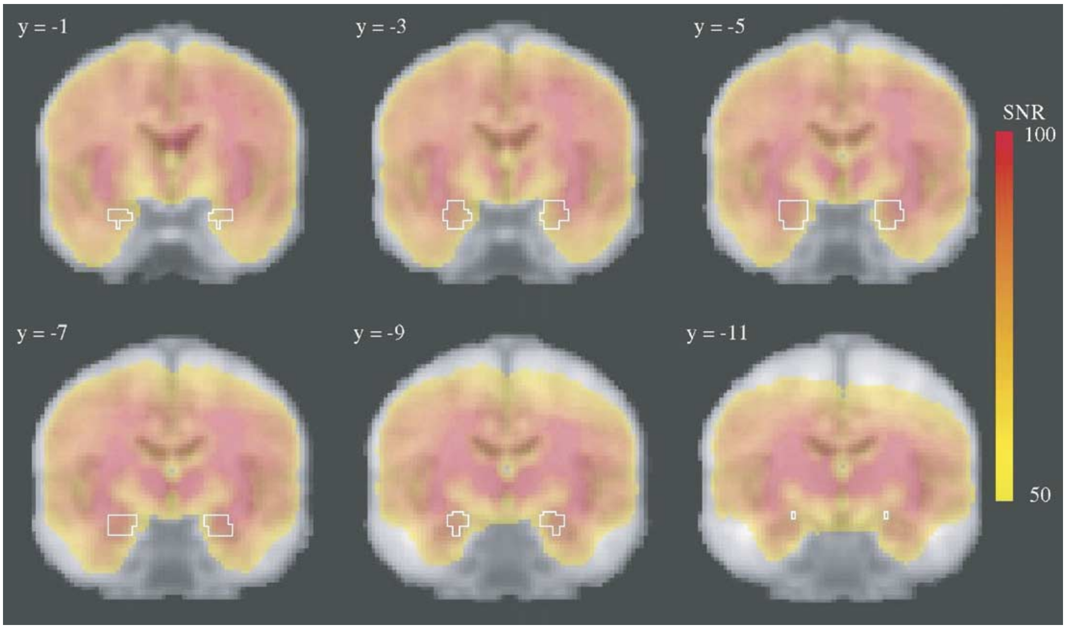

Fig. 2. Two-mm thick coronal sections from $y = -1$ to $y = -11$ depicting mean SNR, with Talairach atlas amygdala ROI outlined. The color scale ranges from an SNR of 50 (yellow) to an SNR of 100 (red).

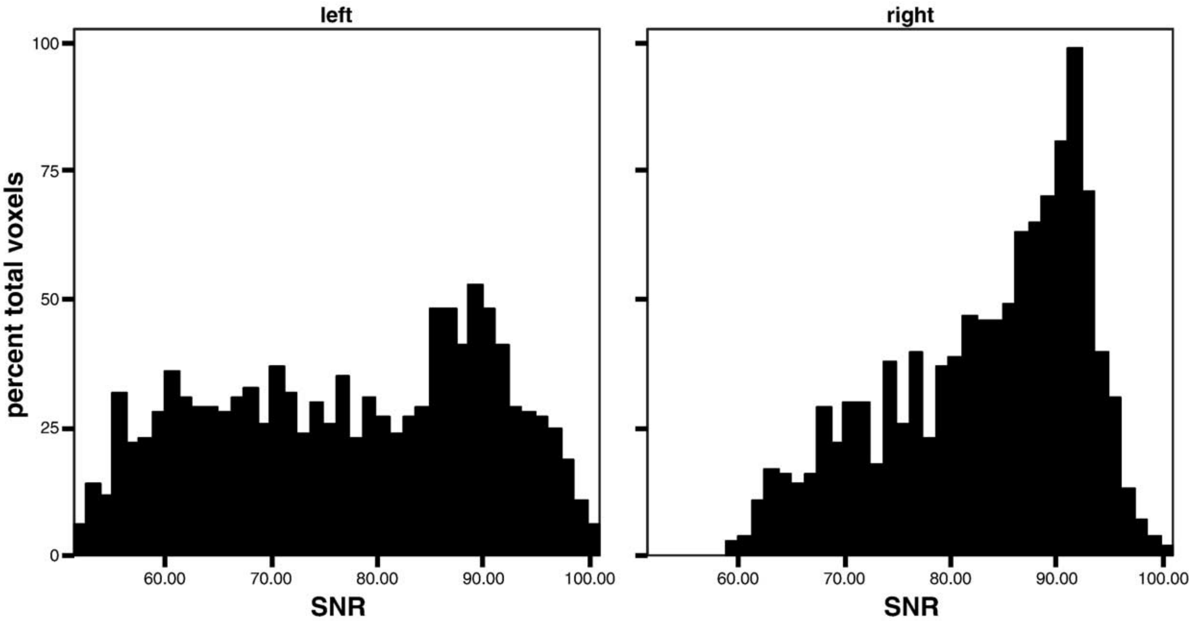

Fig. 3. Histogram of mean voxel SNR values for left and right Talairach-based amygdala ROIs.

quite possibly the low SNR, rather than the lack of an effect per se (LaBar et al., 2001). Given the difficulties in imaging amygdala, in particular medial and ventral amygdala where susceptibility-related dropout is greatest, it would be expedient for future studies to include measures of SNR across the amygdala.

Mixed effects GLM analysis

A voxelwise mixed effects analysis of group data indicated significant bilateral amygdala signal changes across all three scan sessions in response to fearful stimuli, for both blurred and non-blurred z scores and percent signal change (see Fig. 4). The areas of significant activation were almost identical for the percent signal change and z score indices. Mixed effects analysis of mean percent signal change extracted from the Talairach ROI indicated significant bilateral activation for fear versus baseline ($F(1,14) = 18.1$, $P = 0.001$), and fear versus neutral faces ($F(1,14) = 5.67$, $P = 0.032$). These results were identical for the non-blurred and 4 mm blurred data. Mean percent signal change values for the Talairach ROI are shown in the upper section of Table 2. It can be seen that

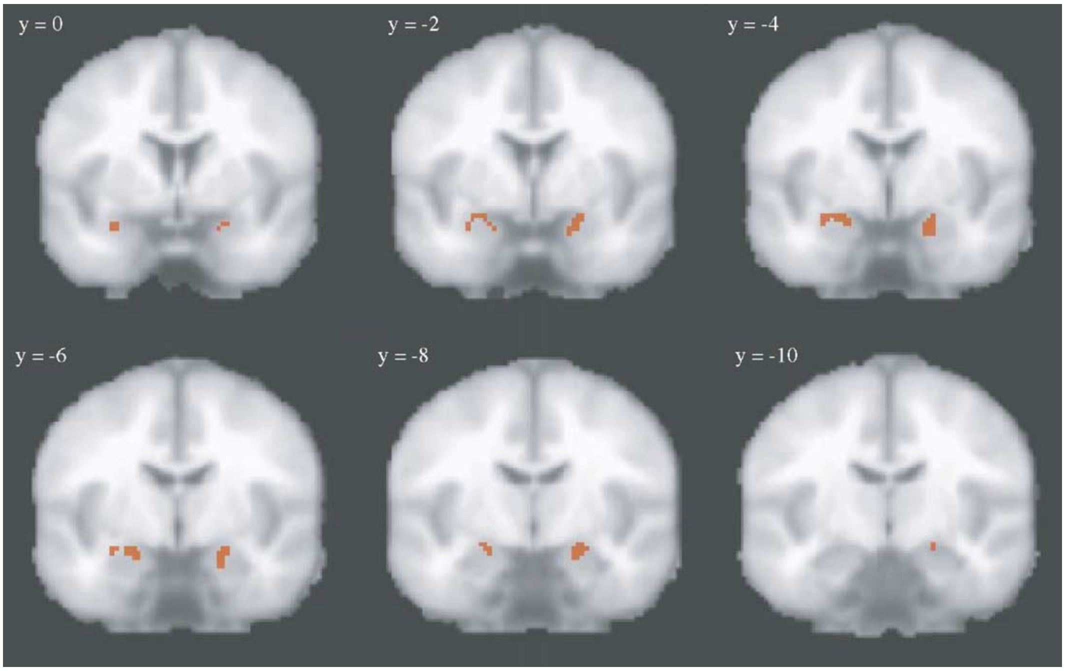

Fig. 4. Statistically defined clusters in left and right amygdala, based upon fear–baseline contrast for scan session 1. This and all other brain images are presented in radiological format (i.e., left of image = right of brain). Images thresholded at $P < 0.01$ corrected for multiple comparisons. Coronal sections are 2 mm thick and extend from $y = 0$ to $y = -10$ (clusters were confined to areas of activation that fell within the amygdala as defined by the Talairach atlas).

Table 2
Contrast indices extracted from Talairach and statistically defined ROIs

	Fear–baseline		Neutral–baseline	
	Left	Right	Left	Right
Percent signal change for Talairach defined ROI				
Session 1	0.17 (0.19)	0.19 (0.19)	0.14 (0.12)	0.13 (0.15)
Session 2	0.15 (0.18)	0.09 (0.22)	0.04 (0.18)	0.02 (0.20)
Session 3	0.16 (0.25)	0.12 (0.26)	0.15 (0.18)	0.11 (0.20)
Percent signal change for statistically defined ROI				
Session 1	0.30 (0.23)	0.24 (0.15)	0.23 (0.13)	0.18 (0.13)
Session 2	0.29 (0.26)	0.14 (0.22)	0.11 (0.17)	0.07 (0.16)
Session 3	0.24 (0.28)	0.15 (0.25)	0.22 (0.16)	0.14 (0.19)
z scores for statistically defined ROI				
Session 1	0.77 (0.64)	0.70 (0.48)	0.70 (0.47)	0.59 (0.47)
Session 2	0.70 (0.59)	0.37 (0.56)	0.32 (0.50)	0.18 (0.52)
Session 3	0.61 (0.58)	0.46 (0.64)	0.74 (0.54)	0.50 (0.57)

Numbers in parentheses are standard deviations.

there is substantial variability across subjects in responses to both fear and neutral faces, with less variability within-subjects across sessions. These values can be put into the context of a one-sample *t* test to test the significance of an overall increase in signal during the condition of interest in a single session (i.e., a positive fear versus baseline or neutral versus baseline contrast). The formula for such a one-sample *t* test is:

$$t = \frac{\overline{X}\sqrt{N}}{s}$$

where $\overline{X}$ is the sample mean, s is the standard deviation, and N is the number of subjects. For a data set with an SNR similar to that reported here (i.e., about 80), and with between-subjects standard deviation of extracted percent signal change approximately equal to the mean signal change, the value of t will equal the square root of the number of subjects. For significance at an alpha of 0.01, this requires approximately 9 subjects; 15 subjects will enable rejection of the null hypothesis at an alpha of 0.001.

A more appropriate way to assess the ability to detect an effect of a given size, given the variability measured here, is to estimate the statistical power, which is equal to 1 minus the probability of falsely accepting the null hypothesis, should a real effect actually exist. If it is assumed that the between-subjects variance measured in this study sample is a good approximation of the population variance in amygdala signal change (given the same acquisition scheme), then one can estimate the statistical power for a given number of subjects and given expected effect size, or conversely how many subjects would need to be included to achieve a given statistical power. Fig. 5 shows the statistical power for varying numbers of subjects and mean contrast values, for the case when the between-subjects standard deviation in signal change is 0.2% or 0.3%. As can be seen, high (i.e., >0.8) statistical power to detect a contrast of 0.2% is achievable with as few as 12–15 subjects with the level of between-subjects variability measured in this experiment. More generally, for acquisition schemes or populations which give rise to greater or lesser between-subjects variability in amygdala contrast estimates, to detect a significant contrast of magnitude equal to the between-subjects standard deviation with a statistical power of 0.8 requires 15 subjects.

It should be noted that these estimates of statistical power pertain to focused statistical tests on specific regions of interest, and do not take into account correction for multiple comparisons in the case of multiple or voxelwise tests. For whole brain voxelwise analyses, it is evident that either (i) a greater signal change, (ii) a lower between-subjects variability in signal change or (iii) a greater number of subjects would be required to exceed corrected statistical thresholds.

Stability of Talairach ROI activation

For the Talairach-defined ROI, intraclass correlation coefficients calculated across the sessions indicated stability of response to fearful faces (as change from baseline and versus response to neutral faces) in left amygdala, but less stability in responses to neutral expressions (see Table 3). In particular, the average measure ICC for fear versus baseline and fear versus neutral contrasts was about 0.50 across all scan sessions, indicating that the mean estimate of fear response measured three times across 8 weeks is a moderately reliable measure. Single measure ICCs of fear response in left amygdala were considerably lower, indicating that mean left amygdala response to fear estimated from individual scan sessions had low reliability. In right amygdala, response to

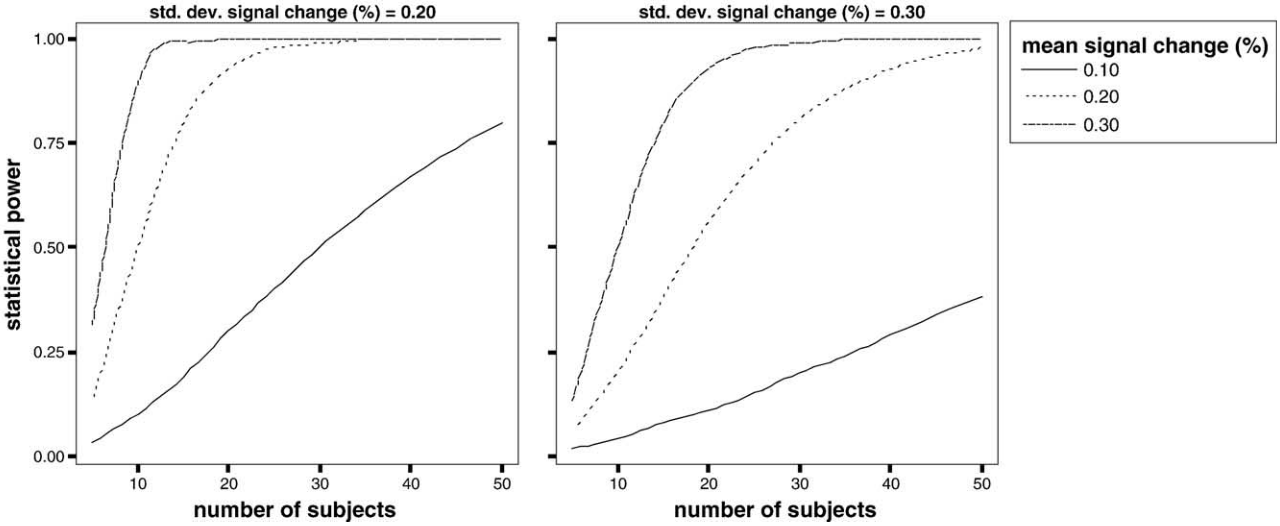

Fig. 5. Plots of statistical power versus number of subjects, given two levels of contrast variability and three hypothesized contrast effect sizes.

Table 3
Intraclass correlation coefficients for the whole, Talairach-defined amygdala

		Across 3 sessions		Pairwise ICCs					
		Left	Right	Left			Right		
				t1–t2	t2–t3	t1–t3	t1–t2	t2–t3	t1–t3
Neutral–fix.	Single	0.08	0.25	−0.21	0.20	0.19	0.22	0.46	−0.01
	Average	0.20	0.50	−0.52	0.34	0.32	0.36	0.63	−0.02
Fear–fix.	Single	0.28	0.18	0.31	0.35	0.19	0.46	0.18	−0.03
	Average	0.53	0.40	0.47	0.51	0.32	0.63	0.31	−0.07
Fear–neutral	Single	026	−0.15	0.24	0.30	0.23	0.28	−0.57	0.00
	Average	0.51	−0.61	0.39	0.46	0.37	0.44	0.00	0.00

fear and neutral expressions showed some stability over 2 weeks, but no stability over longer periods.

Stability of statistically defined ROI activation

Fig. 4 shows the statistically defined clusters in right and left amygdala that showed activation to fear faces at scan session 1. These clusters were then used to extract functional contrast values for the two other scan sessions, as given in the middle section of Table 2. ICCs calculated across sessions for these statistically defined ROIs were considerably higher than for the Talairach ROIs, as shown in Table 4. In particular, the single measure ICC for fear versus baseline in the left amygdala ROI was 0.70 from 0 weeks to 2 weeks, and 0.63 from 0 weeks to 8 weeks, demonstrating a high degree of stability. Single measure fear versus baseline response in the right amygdala ROI was stable over 2 weeks (ICC = 0.55), but not over 8 weeks (ICC = 0.27). The single measure fear versus neutral contrast in the left amygdala ROI showed moderate stability over 2 weeks (ICC = 0.53), but somewhat lower stability over 8 weeks (ICC = 0.42). There was little stability of fear versus neutral in the right amygdala ROI. Unlike response to fear expressions, the neutral versus baseline contrast showed greatest stability in the right amygdala ROI, with single measure ICCs of 0.45 and 0.62 over 2 weeks and 8 weeks, respectively. Stability of neutral versus baseline was low in the left amygdala ROI, which explains why the fear versus neutral contrast in the left amygdala ROI was not as stable as the fear versus baseline measure. Scatterplots across 2 weeks and 8 weeks of the fear versus neutral, fear versus baseline, and neutral versus baseline contrasts, for the left amygdala ROI are presented in Fig. 6. To provide a qualitative indication of whether between-session variability reflected changes to the amplitude, position or the extent of activation, statistical maps of fear versus baseline and neutral versus baseline contrasts for each session are shown in Fig. 7. It can be seen that the extent, location and magnitude of activation for the fear–baseline contrast is highly similar across all three scan sessions. The neutral–baseline contrast, however, shows a large decrease in activated voxels at scan session 2 relative to scan session 1. Activation to neutral faces at scan session 3 was similar to that at scan session 1.

Comparison of blurred and non-blurred images, for z scores and percent signal change

Table 5 shows the single measure ICCs for the left amygdala statistical ROI for 4 mm blurred and non-blurred percent signal change and *z* score data. It is clear that percent signal change provides a somewhat more repeatable measure than *z* scores, although the difference is neither large nor consistent. The effect of blurring with a 4 mm FWHM Gaussian spatial filter was minimal.

Discussion

The current study has demonstrated that it is possible to achieve sufficiently high test–retest reliability in amygdala response to fear faces for such a paradigm to be usefully applied to longitudinal studies. In particular, reliability of left amygdala response to fear faces (compared to baseline) was found to be reliable over a period of 8 weeks. Importantly, however, the contrast of fear faces to neutral faces was not as reliable in left amygdala, mainly due to the unreliability of neutral faces as a comparison condition.

The amygdala has been shown to respond more to the presentation of novel neutral faces than to familiar or repeated neutral faces (Dubois et al., 1999; Wright et al., 2003). Given the reduction in amygdala activation in this study to neutral faces during session 2, it is possible that response to neutral faces diminishes or habituates with familiarity of the stimuli, at least over relatively short time periods. By session 3, responses to neutral stimuli were comparable to session 1, indicating that such

Table 4
Intraclass correlations for the amygdala cluster defined on the basis of fear activation in the first session

		Across 3 sessions		Pairwise ICCs					
		Left	Right	Left			Right		
				t1–t2	t2–t3	t1–t3	t1–t2	t2–t3	t1–t3
Neutral–fix.	Single	0.25	0.57	0.25	0.21	0.33	0.45	0.64	0.62
	Average	0.51	0.80	0.40	0.35	0.50	0.62	0.78	0.77
Fear–fix.	Single	0.67	0.44	0.70	0.68	0.63	0.55	0.50	0.27
	Average	0.86	0.70	0.82	0.81	0.77	0.71	0.67	0.42
Fear–neutral	Single	0.46	0.00	0.53	0.43	0.42	0.37	−0.24	0.02
	Average	0.72	−0.01	0.69	0.60	0.59	0.54	−0.63	0.03

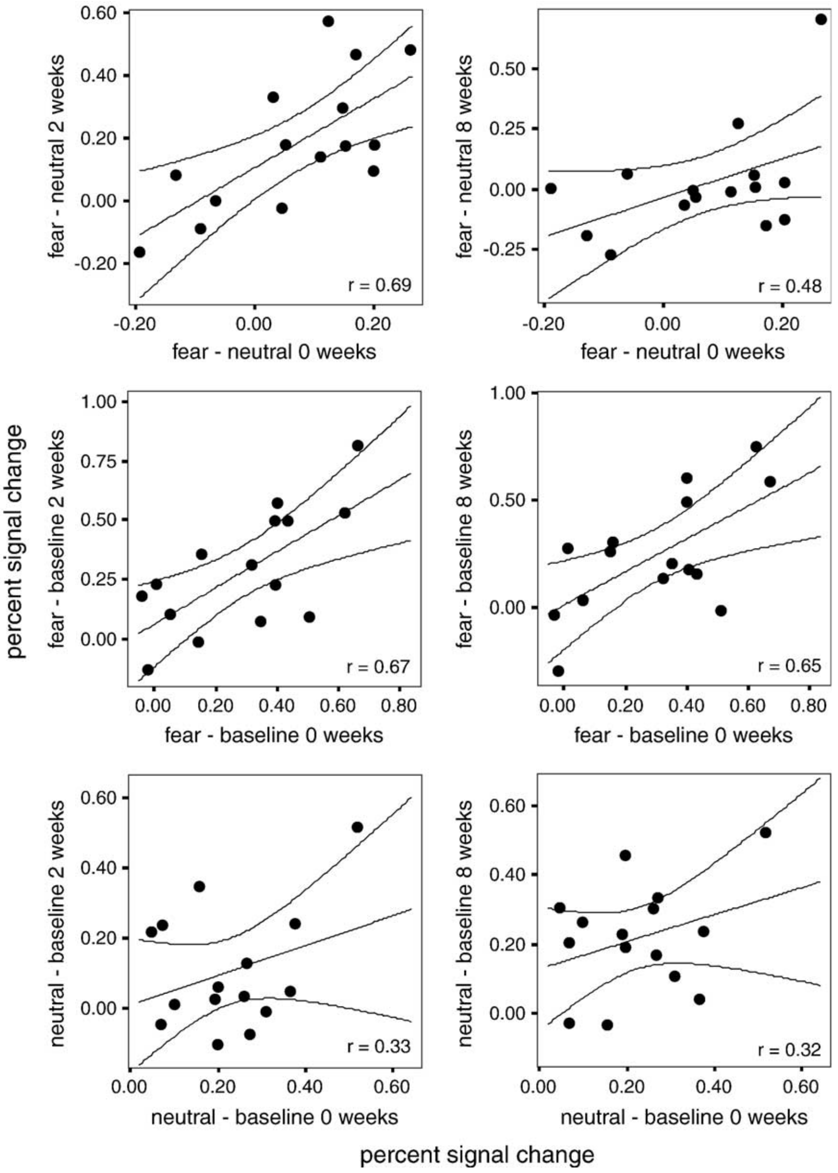

Fig. 6. Scatterplots of fear versus neutral, fear versus baseline and neutral versus baseline for the statistically defined left amygdala cluster. Left column: scatterplots for 0 versus 2 weeks; right column: scatterplots for 0 versus 8 weeks. All values represent percent signal change.

habituation effects might be relatively short-lived, particularly considering that scan sessions 2 and 3 were separated by 6 weeks, as contrasted with the 2 weeks in between scan sessions 1 and 2. If neutral faces are to be usefully employed as the control condition in longitudinal studies of amygdala function, more will need to be known about the extended time course of such habituation effects. It might also be prudent to contrast BOLD response to fear faces either with a simple baseline condition, or perhaps another facial expression that is less emotionally ambiguous than neutral expressions.

We recently demonstrated that ventral amygdala activation to neutral faces correlates with state anxiety, possibly due to the uncertain threat-related value of such facial expressions (Somerville et al., 2004). In this study, however, there was no significant correlation between measures of state anxiety and amygdala activation to neutral faces. This is not surprising, given that only subjects with low anxiety were included in the study, and all subjects reported consistently low anxiety across all three scan sessions. Such a restricted range of anxiety scores makes the finding of significant correlations between anxiety and brain activation unlikely. In other contexts or with different subject groups, it could be expected that variability in state anxiety over periods of weeks could be manifested in increased variability of amygdala response to neutral faces.

Reliability of response to fearful faces was lower in right than in left amygdala. This difference in reliability does not appear to be related to hemispheric differences in signal quality or coverage, since the mean SNR was comparable between the right and left amygdala. Our results are remarkably similar to the reliability of resting state metabolism as measured with PET by Schaefer et al. (2000), who reported a 6 month test–retest ICC of 0.53 in left amygdala and 0.17 in right amygdala. Previous research (Phillips et al., 2001; Wright et al., 2001) has shown greater habituation of right amygdala (compared to left) within session, which might

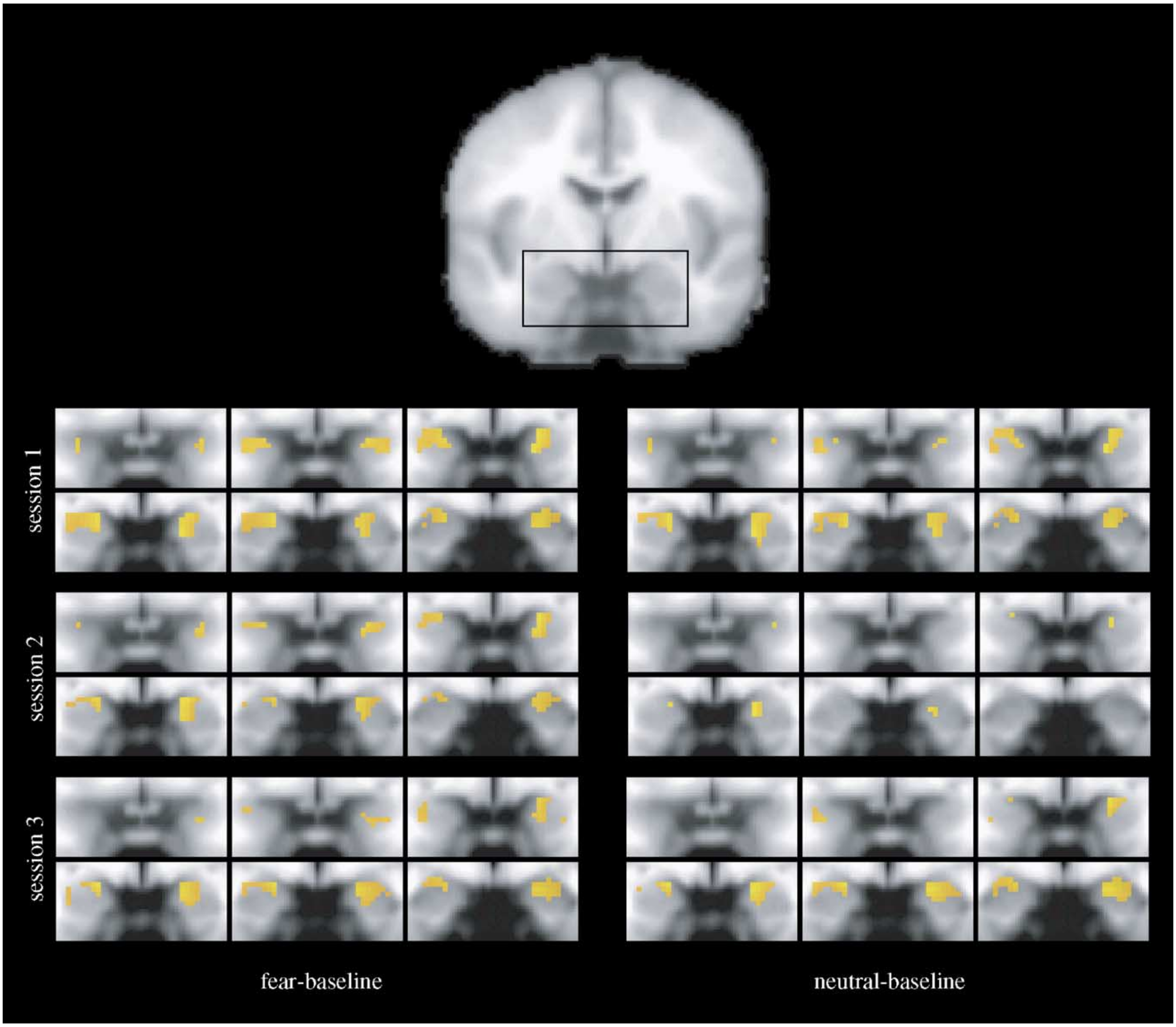

Fig. 7. Statistical maps of fear–baseline (left) and neutral–baseline (right) contrasts for the 3 scan sessions, for the amygdala region as indicated by the rectangle in the whole coronal slice (top). Coronal sections are 2 mm thick and extend from $y = 0$ to $y = -10$. Images thresholded at $P < 0.01$ uncorrected for multiple comparisons. A liberal threshold has been used in these images to permit a comparison of activation across sessions, including activation that would fall just under threshold in a more stringent, corrected statistical test.

decrease right amygdala reliability. Although the current experiment was not designed to permit a formal test of such within-session habituation effects, an examination of average responses to fear or neutral faces over the two scan runs revealed no decrease. To the extent that habituation effects in amygdala response to fearful expressions might exist in other samples, or with other experimental designs, then reliability of amygdala response might be increased if habituation were able to be characterized and/or modeled. The less reliable right amygdala response to fearful faces found here is also potentially consistent

Table 5
Single measure ICCs for the left amygdala statistical ROI for 4 mm blurred and non-blurred percent signal change and *z* score data

		t1–t2		t1–t3	
		% signal change	*z* score	% signal change	*z* score
Neutral–fix.	Non-blurred	0.25	0.00	0.33	0.35
	4 mm blurred	0.29	−0.04	0.42	0.37
Fear–fix.	Non-blurred	0.70	0.57	0.63	0.49
	4 mm blurred	0.69	0.55	0.56	0.48
Fear–neutral	Non-blurred	0.53	0.56	0.42	0.21
	4 mm blurred	0.46	0.57	0.35	0.24

with the suggestion that left amygdala signal changes track the clear categorization or labeling of presented facial expressions (i.e., the current working hypothesis), while right amygdala signal changes are related to the uncertain predictive value of presented facial expressions (i.e., other potential possibilities; Kim et al., 2003).

We also used the present data set to compare the use of percent signal change as an index of fMRI activation with the use of *z* scores. Percent signal change was found to be a slightly more reliable quantification of signal change than *z* scores, although the main contrast effects were very similar. It is possible that in normalizing each subject's signal variation with *z* scores, inter-subject differences are minimized, and thus the somewhat lower test–retest ICCs might reflect a reduction in between-subject range. Although percent signal change would thus seem a better quantification of BOLD signal change, it should be noted that in this study, mean baseline signal was relatively high across the entire amygdala. In cases with greater signal dropout, as might occur when imaging ventral prefrontal cortex, or when imaging amygdala with a less optimal type of acquisition, the use of *z* scores might still be preferable, since they are not directly derived from the mean or baseline signal level.

There may also be conceptual reasons for favoring one measure over the other. Percent signal change is an estimate of the size of signal change, normalized to the amplitude of baseline signal. In contrast, *z* scores are a contrast to noise ratio. The two will yield similar results when the noise scales with the baseline signal. The dominant cause of noise in fMRI experiments is caused by gross subject motion and physiological artifacts, primarily consisting of respiratory and cardiac related motion and susceptibility changes (Jezzard et al., 1993). Such noise might vary across scan sessions as a function of changes in the health or baseline physiological state of an experimental participant. Additionally, it is possible that when comparing different experimental groups of subjects (e.g., controls versus a patient group), the levels of motion-related and physiological noise might be different. In such cases, the *z* scores are likely to be more adversely affected than percent signal change, and might give rise to false between-groups differences in measured activations. In these circumstances, percent signal change would seem a more appropriate measure.

The stability data presented here are not intended to be absolute evidence of the reliability of amygdala response to facial expressions of emotion. The present experimental design must be taken into account. In this study, as with others conducted in our laboratory (see Kim et al., 2003; Somerville et al., 2004), we used a simple block design with passive viewing of the faces. This contrasts with other studies that have used an explicit task, such as identifying the expressed emotion or the sex of the face (e.g., Critchley et al., 2000). It is certainly possible that in this study, without the constraints of an explicit task subjects might have varied with respect to how they attended to the facial expressions. Our decision to avoid having subjects perform a task while viewing facial expressions was based upon evidence that engagement in explicit cognitive or attentional tasks leads to the inhibition of limbic circuitry (Drevets and Raichle, 1998; Shulman et al., 1997; Whalen et al., 1998a,b), and that this modulatory effect is dependent on task engagement and difficulty. According to this view, passive viewing of emotional facial expressions should lead to more robust, and therefore more reliable, amygdala activation than would occur during an experiment in which the subjects engaged in a task. Future studies might compare the reliability and sensitivity of passive versus active tasks in eliciting amygdala activation in response to emotional stimuli.

A further consideration is the experimental context in which fearful facial expressions are presented. In this study, fear expressions were interleaved with happy and neutral expressions. It is likely that amygdala response to both fearful and neutral expressions was influenced by the contrastive presence of happy faces. A number of experiments have demonstrated hedonic contrast effects, whereby stimuli of differing affective valence presented alongside one another, affect reactivity to one another (e.g., Russell and Fehr, 1987). If such contrastive effects are instantiated at the amygdala level, the reliability of fear versus neutral BOLD contrasts would be sensitive to other presented expressions (see Somerville et al., 2004 for discussion of this point). In such types of studies, it might be preferable to use experimental designs in which differently valenced stimuli (e.g., happy versus fearful faces) are presented in separate scan runs with interleaved neutral and fixation control stimuli, with suitable counterbalancing across participants. The collection of behavioral responses to differently valenced stimuli (e.g., subjective ratings, judgement reaction time) would further increase the interpretability of data collected in a similar paradigm. The likely inclusion of both positive and negative valence emotional expressions in future clinical studies (e.g., to measure response to positive facial expressions in anhedonia, depression or social phobia) makes it important to gain a better understanding of such contextual factors and develop methods for lessening their impact on the reliability and generalizability of results.

The likely importance of experimental design and context to the stability of amygdala response to fearful expressions makes the analysis of reliability in control subjects crucial in future longitudinal studies. Given the high cost and ethical considerations involved in clinical research, it would be prudent to examine test–retest reliability of amygdala response in a pilot control group before commencing longitudinal clinical studies. The study design and methods reported here should prove useful for researchers embarking on such research, and provide a benchmark for the level of reliability that should be attainable.

Conclusions

The use of facial expressions of emotion as presented stimuli in uman neuroimaging studies of the amygdala represents a simple and tolerable strategy for assessing potential dysfunction of this system in psychopathology (Rauch et al., 2000; Sheline et al., 2001; Thomas et al., 2001; Yurgelun-Todd et al., 2000). The present study offers information for experimental psychopathologists who might seek to use facial expressions of emotion as a basis for comparing pathological groups with healthy control subjects over three visits to the scanner across 8 weeks. Using the current fMRI acquisition scheme, one can expect reasonable coverage of the amygdaloid region. In addition, responsivity in a healthy control group can be expected to be more reliable within the left, compared to the right, amygdala. Furthermore, comparisons with a fixation baseline will be more stable over time, compared to comparisons with the neutral face condition. Somerville et al. (2004) offer a strategy for measuring state variables that could potentially account for a portion of this variability associated with response to neutral faces. Future studies could determine the relevance of these variations in normal levels of anxiety to understanding amygdala response to neutral faces in pathologically anxious patient groups.

Ackowledgments

We thank Michael Anderle and Ron Fisher for their help in collecting the data. A partial analysis of these data was presented as a poster at the 9th International Conference of the Organization for Human Brain Mapping, New York, 2003. This research was supported by the NIH/NIMH (grants 069315 and 67167) and Wyeth Pharmaceuticals.

References

Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., Buckner, R.L., Strauss, M.M., Hyman, S.E., Rosen, B.R., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17, 875–887.

Chen, N.K., Dickey, C.C., Yoo, S.S., Guttmann, C.R., Panych, L.P., 2003. Selection of voxel size and slice orientation for fMRI in the presence of susceptibility field gradients: application to imaging of the amygdala. NeuroImage 19, 817–825.

Cox, R.W., 1996. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Critchley, H., Daly, E., Phillips, M., Brammer, M., Bullmore, E., Williams, S., Van Amelsvoort, T., Robertson, D., David, A., Murphy, D., 2000. Explicit and implicit neural mechanisms for processing of social information from facial expressions: a functional magnetic resonance imaging study. Hum. Brain Mapp. 9, 93–105.

Drevets, W.C., Raichle, M.E., 1998. Reciprocal suppression of regional cerebral blood flow during emotional versus higher cognitive processes: implication for interactions between emotion and cognition. Cogn. Emot. 12, 353–385.

Dubois, S., Rossion, B., Schiltz, C., Bodart, J.M., Michel, C., Bruyer, R., Crommelinck, M., 1999. Effect of familiarity on the processing of human faces. NeuroImage 9, 278–289.

Ekman, P., Friesen, W.V., 1976. Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto.

Fischer, H., Wright, C.I., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L., 2003. Brain habituation during repeated exposure to fearful and neutral faces: a functional MRI study. Brain Res. Bull. 59, 387–392.

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D.L., Flaum, M., Andreasen, N.C., 1998. Functional MRI statistical software packages: a comparative analysis. Hum. Brain Mapp. 6, 73–84.

Hariri, A.R., Weinberger, D.R., 2003. Functional neuroimaging of genetic variation in serotonergic neurotransmission. Genes Brain Behav. 2, 341–349.

Irwin, W., Davidson, R.J., Lowe, M.J., Mock, B.J., Sorenson, J.A., Turski, P.A., 1996. Human amygdala activation detected with echo-planar functional magnetic resonance imaging. NeuroReport 7, 1765–1769.

Kim, H., Somerville, L.H., Johnstone, T., Alexander, A., Whalen, P.J., 2003. Inverse amygdala and medial prefrontal cortex responses to surprised faces. NeuroReport 14, 2317–2322.

Kim, H., Somerville, L.H., Johnstone, T., Polis, S., Alexander, A.L., Shin, L.M., Whalen, P.J., 2004. Contextual modulation of amygdala responsivity to surprised faces. J. Cogn. Neurosci. 16, 1730–1745.

Jenkinson, M., Bannister, P., Brady, J., Smith, S., 2002. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.

Jezzard, P., Le Bihan, D., Cuenod, C., Pannier, L., Prinster, A., Turner, R., 1993. An investigation of the contribution of physiological noise in human functional MRI studies at 1.5 Tesla and 4 Tesla. 12th Annual Mtg., Proc. Soc. Magn. Reson. Med., pp. 1392.

LaBar, K.S., Gitelman, D.R., Mesulam, M.M., Parrish, T.B., 2001. Impact of signal-to-noise on functional MRI of the human amygdala. NeuroReport 12, 3461–3464.

Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y., Goff, D.C., Rauch, S.L., Kennedy, D.N., Gollub, R.L., 2001. Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158, 955–958.

McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S.J., Holmes, A.P., 2000. Variability in fMRI: an examination of intersession differences. NeuroImage 6, 708–734.

Morris, J., Frith, C., Perrett, D., Rowland, D., Young, A.W., Calder, A.J., Dolan, R.J., 1996. A differential neural response in the human amygdala to fearful and happy facial expressions. Nature 383, 812–815.

Phillips, M.L., Young, A.W., Scott, S.K., Calder, A.J., Andrew, C., Giampietro, V., Williams, S.C., Bullmore, E.T., Brammer, M., Gray, J.A., 1998. Neural responses to facial and vocal expressions of fear and disgust. Proc. R. Soc. Lond., B Biol. Sci. 265, 1809–1817.

Phillips, M.L., Medford, N., Young, A.W., Williams, L., Williams, S.C., Bullmore, E.T., Gray, J.A., Brammer, M.J., 2001. Time courses of left and right amygdalar responses to fearful facial expressions. Hum. Brain Mapp. 12, 193–202.

Rauch, S.L., Whalen, P.J., Shin, L.M., McInerney, S.C., Orr, S., Lasklo, N., Pitman, R., 2000. Exaggerated amygdala response to masked facial expressions in posttraumatic stress disorder. Biol. Psychiatry 47, 769–776.

Robinson, S., Windischberger, C., Rauscher, A., Moser, E., 2004. Optimized 3 T EPI of the amygdalae. NeuroImage 22, 203–210.

Russell, J.A., Fehr, B., 1987. Relativity in the perception of emotion in facial expressions. J. Exp. Psychol. Gen. 116, 223–237.

Schaefer, S.M., Abercrombie, H.C., Lindgren, K.A., Larson, C.L., Ward, R.T., Oakes, T.R., Holden, J.E., Perlman, S.B., Turski, P.A., Davidson, R.J., 2000. Six-month test–retest reliability of MRI-defined PET measures of regional cerebral glucose metabolic rate in selected subcortical structures. Hum. Brain Mapp. 10, 1–9.

Schwartz, C.E., Rauch, S.L., 2004. Temperament and its implications for neuroimaging of anxiety disorders. CNS Spectr. 9, 284–291.

Sheline, Y.I., Barch, D.M., Donnelly, J.M., Ollinger, J.M., Snyder, A.Z., Mintun, M.A., 2001. Increased amygdala response to masked emotional faces in depressed subjects resolves with antidepressant treatment: an fMRI study. Biol. Psychiatry 50, 651–658.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 2, 420–428.

Shulman, G.L., Fiez, J.A., Corbetta, M., Buckner, R.L., Miezin, F.M., Raichle, M.E., et al., 1997. Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. J. Cogn. Neurosci. 9, 648–663.

Smith, S., Beckmann, C., Ramnani, N., Woolrich, M., Bannister, P., Jenkinson, M., Matthews, P., McGonigle, D., in press. Variability in fMRI: a re-examination of intersession differences. Hum. Brain Mapp.

Somerville, L.H., Kim, H., Johnstone, T., Alexander, A., Whalen, P.J., 2004. Human amygdala response during presentation of happy and neutral faces: correlations with state anxiety. Biol. Psychiatry 55, 897–903.

Talairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme, New York.

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

Thomas, K.M., Drevets, W.C., Dahl, R.E., Ryan, N.D., Birmaher, B., Eccard, C.H., Axelson, D., Whalen, P.J., Casey, B.J., 2001. Amygdala response to fearful faces in anxious and depressed children. Arch. Gen. Psychiatry 58, 1057–1063.

Whalen, P.J., Bush, G., McNally, R.J., Wilhelm, S., McInerney, S.C., Jenike, M.A., Rauch, S.L., 1998a. The emotional counting Stroop paradigm: a functional magnetic resonance imaging probe of the anterior cingulate affective division. Biol. Psychiatry 44, 1219–1228.

Whalen, P.J., Rauch, S.L., Etcoff, N.L., McInerney, S.C., Lee, M.B., Jenike, M.A., 1998b. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. J. Neurosci. 18, 411–418.

Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Wright, C.I., Rauch, S.L., 2001. A functional MRI study of human amygdala responses to facial expressions of fear versus anger. Emotion 1, 70–83.

Wright, C.I., Fischer, H., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L., 2001. Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. NeuroReport 12, 379–383.

Wright, C.I., Martis, B., Schwartz, C.E., Shin, L.M., Fischer, H., McMullin, K., Rauch, S.L., 2003. Novelty responses and differential effects of order in the amygdala, substantia innominata, and inferior temporal cortex. NeuroImage 18, 660–669.

Yurgelun-Todd, D.A., Gruber, S.A., Kanayama, G., Killgore, W.D., Baird, A.A., Young, A.D., 2000. fMRI during affect discrimination in bipolar affective disorder. Bipolar Disord. 2, 237–248.