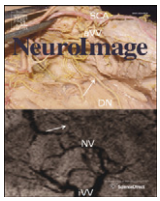
NeuroImage 60 (2012) 1746–1758

Contents lists available at SciVerse ScienceDirect

NeuroImage

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



Full Length Article

Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery

Michael M. Plichta [a,d,*], Adam J. Schwarz [b], Oliver Grimm [a], Katrin Morgen [a], Daniela Mier [c], Leila Haddad [a], Antje B.M. Gerdes [d], Carina Sauer [c], Heike Tost [a], Christine Esslinger [a,e], Peter Colman [f], Frederick Wilson [g], Peter Kirsch [c], Andreas Meyer-Lindenberg [a]

[a] Central Institute of Mental Health, Department of Psychiatry and Psychotherapy, University of Heidelberg/Medical Faculty Mannheim, Mannheim, Germany
[b] Eli Lilly and Company, Translational Medicine, Indianapolis, USA
[c] Central Institute of Mental Health, Department of Clinical Psychology, University of Heidelberg/Medical Faculty Mannheim, Mannheim, Germany
[d] Chair of Clinical and Biological Psychology, School of Social Sciences, University of Mannheim; Germany
[e] Department of Neurology, University Hospital of Magdeburg, Germany
[f] Pfizer Worldwide Research and Development, Research Statistics, Sandwich, UK
[g] Pfizer Worldwide Research and Development, Pharmatherapeutics Precision Medicine, Sandwich, UK

ARTICLE INFO

Article history:
Received 4 November 2011
Revised 26 January 2012
Accepted 28 January 2012
Available online 8 February 2012

Keywords:
fMRI
Reliability
Reproducibility
Intra-class correlation coefficient
Working memory
Emotion
Reward

ABSTRACT

Even more than in cognitive research applications, moving fMRI to the clinic and the drug development process requires the generation of stable and reliable signal changes. The performance characteristics of the fMRI paradigm constrain experimental power and may require different study designs (e.g., crossover vs. parallel groups), yet fMRI reliability characteristics can be strongly dependent on the nature of the fMRI task. The present study investigated both within-subject and group-level reliability of a combined three-task fMRI battery targeting three systems of wide applicability in clinical and cognitive neuroscience: an emotional (face matching), a motivational (monetary reward anticipation) and a cognitive (n-back working memory) task. A group of 25 young, healthy volunteers were scanned twice on a 3 T MRI scanner with a mean test–retest interval of 14.6 days. FMRI reliability was quantified using the intraclass correlation coefficient (ICC) applied at three different levels ranging from a global to a localized and fine spatial scale: (1) reliability of group-level activation maps over the whole brain and within targeted regions of interest (ROIs); (2) within-subject reliability of ROI-mean amplitudes and (3) within-subject reliability of individual voxels in the target ROIs. Results showed robust evoked activation of all three tasks in their respective target regions (emotional task = amygdala; motivational task = ventral striatum; cognitive task = right dorsolateral prefrontal cortex and parietal cortices) with high effect sizes (ES) of ROI-mean summary values (ES = 1.11–1.44 for the faces task, 0.96–1.43 for the reward task, 0.83–2.58 for the n-back task). Reliability of group level activation was excellent for all three tasks with ICCs of 0.89–0.98 at the whole brain level and 0.66–0.97 within target ROIs. Within-subject reliability of ROI-mean amplitudes across sessions was fair to good for the reward task (ICCs = 0.56–0.62) and, dependent on the particular ROI, also fair-to-good for the n-back task (ICCs = 0.44–0.57) but lower for the faces task (ICC = −0.02–0.16). In conclusion, all three tasks are well suited to between-subject designs, including imaging genetics. When specific recommendations are followed, the n-back and reward task are also suited for within-subject designs, including pharmaco-fMRI. The present study provides task-specific fMRI reliability performance measures that will inform the optimal use, powering and design of fMRI studies using comparable tasks.

Introduction

There is increasing interest in the potential application of fMRI as an imaging biomarker to probe therapeutic interventions, individualize therapy, and provide proof of concept (Barch and Mathalon, 2011; Borsook et al., 2006; Patin and Hurlemann, 2011; Schwarz et al., 2011a, 2011b; Wise and Preston, 2010; Wise and Tracey, 2006), potentially combined with specific genotypes as an imaging intermediate phenotype (Meyer-Lindenberg, 2010; Meyer-Lindenberg and Weinberger, 2006). For prospective use in these contexts and to maximize confidence in the results, the ability of fMRI paradigms to generate a stable and reliable signal change amenable to modulation by the chosen intervention and experimental design is paramount. Because fMRI reliability characteristics can be strongly dependent on the particular fMRI paradigm employed (Bennett and Miller,

* Corresponding author at: Central Institute of Mental Health, Department of Psychiatry and Psychotherapy, University of Heidelberg/Medical Faculty Mannheim, J5, 68159 Mannheim, Germany. Fax: +49 621 1703 706501.
E-mail address: Michael.Plichta@zi-mannheim.de (M.M. Plichta).

1053-8119 © 2012 Elsevier Inc. Open access under CC BY-NC-ND license.
doi:10.1016/j.neuroimage.2012.01.129

2010), it is critical to formally examine the reliability measures for the specific fMRI tasks to be applied. From a technical point of view, fMRI scans are nowadays often performed with higher field strengths compared to earlier investigations. Therefore, reliability characteristics determined at e.g. 1.5 Tesla (T) may not generalize to 3 T. With regards to study designs, knowledge of the within-subject and group-level reliability of a given paradigm will contribute to how an intervention study should best be arranged (e.g., crossover vs. parallel groups).

Previous fMRI test–retest studies have quantified fMRI reliability for a range of paradigms, from basal sensory stimulation to complex cognitive tasks (Caceres et al., 2009; Gountouna et al., 2010; Lee et al., 2010; Liou et al., 2003; Machielsen et al., 2000; Maiza et al., 2010; Manoach et al., 2001; Miki et al., 2001; Rombouts et al., 1997, 1998; Specht et al., 2003; Stark et al., 2004; Tegeler et al., 1999; Wagner et al., 2005; Wei et al., 2004; Yetkin et al., 1996). These studies converge to the conclusion that group activation maps are highly reproducible across measurement sessions and across different scanners, whereas single subject amplitudes are less reliable. A recent overview (Bennett and Miller, 2010) reported the mean intraclass-correlation coefficient (ICC) of 13 studies on intra-subject BOLD amplitude reliability to be ICC = 0.50 but with a large variance across different studies (ICC = 0.16–0.88). The ICC is a widely used reliability index (Shrout and Fleiss, 1979) ranging from 0 (unreliable) to 1 (perfect reliability)[1]. Factors contributing to the large variance in ICCs might be fMRI scanner-specific (magnet strength; scanner stability, signal-to-noise ratio); sample- (cohort size and composition) and subject-specific (cognitive state across subjects and time; task comprehension) or task-specific (intra- and inter-session habituation/training-effects; blocked versus event-related designs; target region size).

The present study focused on task-specific effects while the other factors (scanner and sample) were held constant or were controlled for as well as possible. Specifically, we investigated the within-subject and group-level reliability of three fMRI tasks in the same group of young, healthy subjects. The fMRI test battery was constructed to cover three fundamental dimensions of human information processing—emotional, motivational and cognitive domains—within a reasonable time span of one scan session. The particular tasks were selected due to their time-efficiency and their widespread use in prior studies suggesting robust performance. Furthermore, these tasks cover a wide range of complementary hypothesized deficits in psychiatric diseases. The task battery consisted of (1) an emotional face matching task (Hariri et al., 2002) that evokes bilateral BOLD signal increases in the amygdala among other regions. The amygdala signal has been shown to be sensitive to genetic variants linked to depression, anxiety, aggression and neuroticism (Meyer-Lindenberg et al., 2006; Pezawas et al., 2005); (2) a reward paradigm (Kirsch et al., 2003) that evokes signal in the ventral striatum/nucleus accumbens (VS/NAcc) which has been found to be sensitive to genetic variants (Forbes et al., 2007; Hahn et al., 2011; Kirsch et al., 2006) and is linked to impulsivity and (an-)hedonic states in clinical and non-clinical populations (Forbes et al., 2010; Hahn et al., 2009; Kirsch et al., 2006; Plichta et al., 2009; Scheres et al., 2007) and (3) an n-back working memory paradigm (Callicott et al., 1998) evoking BOLD signal increases within the right DLPFC and bilateral parietal cortices. Response to this task has been shown to be abnormal in schizophrenia (Glahn et al., 2005), related to heritable risk and sensitive to genetic variation in candidate and genome-wide significant variants for the disorder genes (Esslinger et al., 2009; Meyer-Lindenberg et al., 2007; Tan et al., 2007).

Previous reports of reliability in similar paradigms provide important context for the present work. For emotional processing, Johnstone et al. (2005) examined the amygdala BOLD response to fearful faces contrasted against both neutral faces and fixation-cross over three scan sessions and reported single-measure ICCs for the anatomically defined amygdala ROIs of 0.30 (two-week test interval). Considerably higher single measure ICCs were obtained (0.53 for fearful vs. neutral and 0.70 for fearful vs. fixation) when post-hoc statistically defined amygdala ROIs, based on significant session #1 activation clusters, were used, indicating that the strongest responding voxels provided the most reliable signal. Another study by Schacher et al. (2006) used visual presentations of dynamic fearful faces presented in a block-design and found high amygdala activation reliability (ICC = 0.69–0.83). With an emotional scene paradigm also targeting the amygdala, Stark et al. (2004) reported large changes in BOLD response across sessions and in general low similarity of the test–retest signals (median of Cohen's Kappa < 0.1) including the amygdala. Finally, Manuck et al. (2007) report long-term reliability with a retest interval > 1 year of the emotional face task in a range of ICC = 0.59 for the right amygdala but not for left amygdala (ICC = −0.08). In a study of reward processing, Fliessbach et al. (2010) quantified the reliability of three different reward tasks (all of an event-related design). Dependent on the particular reward task variant, contrast and hemisphere, they reported ventral-striatal ROI ICCs of −0.15 to 0.44 with a mean ICC of < 0.1. Using a working-memory paradigm, Caceres et al. (2009) reported ICCs based on the median subject-level contrast values within ROIs in the right dorsolateral prefrontal cortex (rDLPFC) of 0.44 and in the parietal cortex of 0.55 (left) and 0.36 (right). Finally, a study on the heritability of working memory brain activation reported voxel-wise ICC in most activated areas of 0.7–0.9 (Blokland et al., 2011).

However, the above studies vary with regard to the paradigm, sample size (N = 10–40), magnet strength (1.5 T, 3 T and 4 T), test–retest interval and scanner parameters. All of these factors may have had an impact on the estimated reliability. Therefore, the objective of the present study was to examine the reliability of three distinct tasks in the same sample (N = 25) of healthy subjects while factors potentially impacting the reliability were held constant or controlled.

To comprehensively evaluate the test–retest performance of the three tasks, reliability was quantified at three levels ranging from a broad and global to a localized and fine grained scale: (1) reliability of group-level activation maps over the whole brain and within targeted ROIs (Raemaekers et al., 2007; Specht et al., 2003); (2) within-subject reliability of ROI-mean amplitudes and (3) characteristics of the voxel-scale, within-subject ICCs within the ROIs (Caceres et al., 2009). We also quantified effect sizes and components of variance for the task battery. Together, these performance measures will inform the optimal use, powering and design of fMRI studies using these tasks.

Methods

Subjects

Twenty-five healthy subjects (10 males) were scanned twice (mean retest interval was 14.6 days, S.D. 2.1, range 12–21), while they performed three tasks presented in a fixed order (n-back, faces, reward). Because the intended maximum retest interval of 21 days was exceeded, three additional subjects were not included in analysis. We only included right-handed subjects. The mean age was 24.4 (S.D. 2.8, range 20–32). Further exclusion criteria were positive screening of DSM-IV axis I and II disorders, history of neurological disorders, and regular use of any medication. We assessed hours of sleep, cigarettes smoked and caffeine intake (cups of coffee/caffeinated tea) before beginning the first fMRI session and provided this information to the

[1] Under some conditions negative ICCs can emerge, implying negative reliability. This is theoretically difficult to interpret (Rousson et al., 2002) and the reasons for negative ICC values are not completely understood (Muller and Buttner, 1994). In the present study negative ICC are reported and interpreted as mirroring complete unreliability (= zero).

subject as a reminder for the second session—subjects were asked to come in a comparable state with regards to these measures (all *p*-values > 0.10—see Supplementary Table 1).

All participants were informed about the nature of the experiment as well as the operating mode of the MRI scanner before providing written informed consent. The fMRI investigation of healthy participants and the whole experimental procedure was in accordance with the Declaration of Helsinki and was approved by the local ethics committee of the medical faculty Mannheim of the University of Heidelberg.

Paradigms

The faces task (Hariri et al., 2002) targets emotional processing and is designed to activate the amygdala bilaterally. Subjects viewed a sequence of either fearful or angry faces (experimental condition) or geometric forms (control condition), in alternating blocks of ~30 s each (each trial per block was presented for 5 s). In each condition, each visual presentation comprises three pictures, one (the target image) centered at the top above two test images positioned left and right at the bottom. One of the test images is identical to the target image and the subject must identify it by a left or right button press. Four blocks are presented for each condition, with a total run length of 4 min 28 s (Table 1).

The reward task (Kirsch et al., 2003) targets the reward system and is designed to robustly activate the ventral striatum (VS) including the nucleus accumbens (NAcc). The subject must respond sufficiently quickly to a light-flash on the visual display screen. The flash is preceded by an arrow icon that informs the subject about the consequences of their response to the flash stimulus. Four conditions are included in the paradigm: (1) win condition (arrow up): the subject will win 2 Euros if the response is sufficiently fast; (2) avoidance of loss condition (arrow down): the subject will lose 2 Euros if the response is too slow; (3) verbal control (vertical double arrow): only written feedback is given (no gain or loss of money); (4) passive control condition (horizontal double arrow): no response required. This is an event-related paradigm, in which each of the above conditions is presented 10 times in a pseudo randomized order. The reaction time window is adaptively tailored to the individual response times of the subject in order to have comparable winnings across subjects. The total run length was 8 min 54 s (Table 1).

The n-back task (Callicott et al., 1998) is a working memory paradigm designed to activate the dorsolateral prefrontal cortex (DLPFC), usually predominantly on the right, and the parietal cortices bilaterally. Subjects viewed a series of digits (1–4) presented sequentially for 500 ms (inter-stimulus interval = 1500 ms). One of the numbers in each frame is highlighted and represents the target number to be maintained in memory. As the sequence progresses, the subject must indicate via a button press the highlighted number corresponding either to the currently displayed frame (0-back, control condition) or two frames previously (2-back, experimental condition). The stimuli are presented in a block design; each block lasts 28 s and four blocks are presented for each condition. The conditions are alternated, and the total run length is 4 min 16 s (Table 1).

All paradigms were presented to the subject via LCD video goggles controlled by the software Presentation©. To allow familiarization with the equipment and tasks, participants were carefully instructed and performed short test versions of the paradigms outside the scanner. Because of its relative difficulty, the working-memory task was trained more intensively, i.e. subjects trained the task until more than 60% correct responses during 2-back condition were achieved.

Image acquisition

All MRI sequences were performed on a 3.0-Tesla whole body scanner (Magnetom Trio, Siemens Medical Solutions, Erlangen, Germany). Prior to the functional images, a high-resolution T1-weighted 3D MRI sequence was conducted (ascending slices with a slice thickness = 1.0 mm, FOV = 256 mm × 256 mm × 256 mm, matrix = 256 × 256 × 256). For each paradigm, identical coverage of the whole brain was used including cerebellum, scalp, eyes and nose to avoid wrap-around artifacts. For all paradigms and across both sessions, functional data was acquired using identical echo planar imaging (EPI) sequences with the following scanning-parameters: TR/TE = 2000/30 ms; flip angle = 80°; 28 axial slices (slice-thickness = 4 mm + 1 mm gap) ascending, FOV = 192 mm × 192 mm × 192 mm, matrix = 64 × 64 × 64.

Scanner quality assurance (QA)

Quality assurance (QA) measures were conducted on every measurement day according to an established QA protocol (Friedman and Glover, 2006) quantifying scanner magnet stability using a phantom. The QA protocol includes the following metrics: mean signal intensity (MS), spatial signal-to-noise ratio (sSNR), temporal signal-to-noise ratio (tSNR), signal-to-fluctuation-noise ratio (SFNR), percent signal fluctuation (%Fluct) and percent signal drift (%Dft). A standard water filled cylindric plastic bottle phantom (1900 ml water with 7.125 g $NiSO_4$ and 9.5 g NaCl) provided by the manufacturer (Siemens Medical Systems, Erlangen, Germany) was placed in the epicenter of the scanner, 150 volumes were acquired, using the same T2* weighted EPI sequence that was used for scanning the participants (see above for sequence details). All QA metrics were stable across sessions (all p-values > 0.10; see Supplementary Table 2).

fMRI data quality control (QC)

Quality control (QC) measures of all fMRI time series were integrated into the processing pipeline. The QC protocol included the following metrics: maximum translational excursion calculated from the root mean square (RMS) of the three translational motion vectors, maximum translational excursion calculated from the root mean square (RMS) of the three rotational vectors transformed into translations at the brain edge by the relation $d = r\theta$ with $r = 85$ mm approximating the antero-posterior head radius, the sum of the volume-to-volume translational excursions through the time series, the sum of the volume-to-volume rotational excursions through the time series, the sum of the absolute value of the volume-to-volume translational excursions through the time series, and the sum of the absolute value of the volume-to-volume rotational excursions through the time series.

Table 1
Task characteristics.

	Faces	Reward	n-back
Task duration	4:28 min	8:54 min	4:16 min
Task design	Blocked	Event-related	Blocked
Regressors of interest	Faces; forms	Win; verbal; loose; neutral	2-back; 0-back
Additional regressors	6 movement parameters + constant	6 movement parameters + constant	6 movement parameters + constant
High-pass filter (Hz)	1/128	1/128	1/128
Second-level contrasts of interest	Faces > forms	Win > verbal	2-back > 0-back
Target structure	Amygdala	Ventral-striatum/nucleus accumbens	Right DLPFC + parietal cortex

These QC measures revealed very stable time series across subjects, sessions and tasks with excursions substantially less than the size of a functional voxel. Results are shown in Supplementary Table 3.

fMRI data analyses

The fMRI data were analyzed using statistical parametric mapping (SPM8; Wellcome Department of Cognitive Neurology, Institute of Neurology, London, United Kingdom). Preprocessing of the fMRI data for all three tasks was identical (except for slice-time correction) and included motion correction, spatial normalization into Montreal Neurological Institute [MNI] space and resampling to $2\times2\times2$ mm^3, and spatial smoothing with an 8-mm full-width at half maximum (FWHM) Gaussian kernel. For the reward task, which is an event-related design, we additionally performed slice-time correction prior to motion correction. Spatial normalization was performed by calculating linear (12-parameter affine) and nonlinear transformations of the mean EPI image from each time series to the SPM EPI template in MNI space, and then applying these same transformation parameters to the time series. We additionally ran all analyses with indirect normalization, i.e. high resolution T1 images from session #1 were co-registered to the mean EPI image. The T1 image is then normalized to MNI space (via SPM procedure "segment") and the normalization parameters are then applied to all EPI images.

Statistical analyses comprised first level temporal modeling within a general linear model (GLM) framework to generate a 3D map corresponding to estimated regressor response amplitudes. A complete list of regressors corresponding to the task specific design matrices is presented in Table 1. Regressors of interest were convolved with the default SPM hemodynamic response function (HRF) computed as a 2-parameter gamma function. Motion parameters were not convolved with the HRF. For all three tasks a high-pass filter with a cut-off frequency of 1/128 Hz was used to attenuate low frequency components. All analyses were corrected for serial correlated errors by fitting a first-order autoregressive process (AR[1]) to the error term.

To obtain fMRI group level effects, the particular 25 contrast-of-interest images served as input data for second-level one-sample t-tests. For all tasks, the significance threshold for group-level contrasts was set to $p<.05$, family-wise error (FWE) corrected for multiple comparisons within the pre-specified regions of interest, based on Gaussian Random Field theory.

Regions-of-interest (ROI) definitions

For the faces task the ROI mask "amygdala" was taken from the WFU-PickAtlas (Version 2.5, Wake Forest University, School of Medicine, Winston-Salem, North Carolina; www.ansir.wfubmc.edu), atlas = "human-atlas aal", and left and right amygdalae were treated as separate ROIs. For the reward task, the ventral-striatum (VS) was a fusion of mask "caudate head" taken from WFU-PickAtlas (human-atlas TD brodmann areas +) and mask "accumbens" from the Harvard–Oxford Subcortical Structural Atlas (implemented in FSLView 3.1.8; see http://www.cma.mgh.harvard.edu/fsl_atlas.html; probability threshold was set to 50%) and left and right VS were treated as separate ROIs. For the n-back task, we used empirical masks based on binarized second-level activation maps (2-back > 0-back) calculated from an independent subject sample that was scanned using the same paradigm ($n=60$)—see Supplementary material for more details. This resulted in five ROIs reflecting brain regions strongly responding to the task: two in the right dorsolateral prefrontal cortex and three in the parietal cortex—one left, one right and one medial (these empirical masks are available on request). Atlas structures covering these parts of the brain were larger than typically observed activation foci, hence probably heterogeneous in function and likely to result in low ROI mean response values due to a mixing of strongly and weakly responding voxels. The ROI masks for all three tasks are shown in Fig. 1.

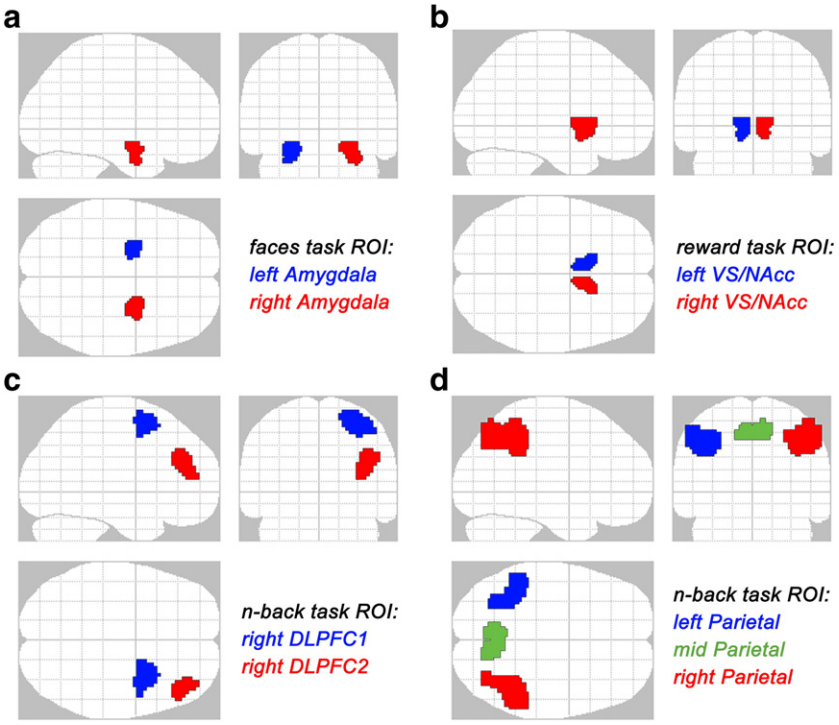


Fig. 1. ROI definitions for the three tasks: (a) shows ROIs for the faces task, i.e., left and right amygdala mask; (b) ROIs for the reward task, i.e. left and right ventral striatum including nucleus accumbens (VS/NAcc); empirical ROIs for the n-back task, including (c) two right DLPFC (1 and 2) definitions and (d) three parietal cortex ROIs. For details, see Methods section.

Statistical methods—reliability

To comprehensively evaluate the test–retest reliability of the three tasks, we examined both the group-level consistency of the fMRI responses and the within-subject reliability across sessions. In addition, we evaluated the stability of the group-mean responses from the first session only, to guide the use of these tasks in parallel group designs. Furthermore, reliability of the recorded behavioral data was also analyzed.

Reliability was assessed using two variants of the ICC, namely ICC(2,1) and ICC(3,1), defined by Shrout and Fleiss (1979) as:

$$ICC(2,1) = BMS - EMS / (BMS + (k-1) * EMS + k * (JMS - EMS)/N) \quad (1)$$

$$ICC(3,1) = BMS - EMS / BMS + (k-1) * EMS \quad (2)$$

where BMS = between-subjects mean square; EMS = error mean square; JMS = session mean square (the original terminology of "J" is "Judge"); k = number of repeated sessions and n = number of subjects. Thus, in the current study, k = 2 and n = 25.

The calculation of both these variants allowed us to determine the reliability in terms of relative (consistent measures = ICC(3,1)) or absolute agreement (ICC(2,1)). Both forms of the ICC estimate the correlation of the BOLD fMRI signal intensities between sessions, modeled by a two-way ANOVA. In the case of ICC(2,1), both effects (subjects and sessions) are assumed to be random, while for ICC(3,1) the effect of sessions is assumed to be fixed. Following Fleiss (1986), we denote ICC values <0.4 as poor, 0.4–0.75 as fair to good and >0.75 as excellent.

Analyses were done using PASW Statistics 18 (IBM SPSS Statistics; Chicago, IL) and MATLAB 7.7 (The Mathworks, Natick, MA).

Consistency of the group-level fMRI responses across sessions

Effect sizes (ES) for each session were calculated at both voxel and ROI level. We report ES as the mean BOLD response divided by the standard deviation across subjects, from each session independently. We also performed paired t-tests between the ROI-mean responses in each session to assess systematic bias between the first and second session. Furthermore, the spatial overlap between group activation maps (Rombouts et al., 1997, 1998) was calculated for each task and its respective ROIs:

$$R_{OVERLAP} = 2 * A_{OVERLAP} / (A1 + A2) \quad (3)$$

where A1 and A2 represent the quantity of the activated voxels of the first and second session, respectively. $A_{OVERLAP}$ is the quantity of identical supra-threshold voxels in both sessions. $R_{OVERLAP}$ ranges from 0 (worst) to 1 (best) or can be expressed as a percentage.

To test the consistency of the group-level spatial distribution of the BOLD signal independent of a statistical threshold, all second-level contrast values from session #2 were plotted against those from session #1 (Raemaekers et al., 2007; Specht et al., 2003). This was done for all voxels within the whole brain and within the target ROIs. The reliability of these group-level changes was quantified using R^2 (i.e., coefficient of determination) and both ICC variants.

Within-subject reliability

The reliability of the BOLD responses within subjects was assessed using the ICCs (1) from the ROI-mean amplitudes (mean contrast value across all voxels in the ROI from each subject and session) and (2) from the contrast amplitudes of each voxel in the ROI (for each subject and session), leading to a distribution of voxel-scale ICC values for each target region (Caceres et al., 2009).

Bland–Altman plots

Within subject reliability of the ROI-mean amplitudes was also evaluated graphically by ladder and Bland–Altman plots (Bland and Altman, 1986). The ladder plots track each subject's BOLD signal change across both sessions and enable a visual assessment of the reproducibility within subjects. The Bland–Altman plots depict the difference versus the mean of the measures from the two sessions and serves as a visual check that the magnitude of the differences is comparable throughout the range of measurement. These calculations were performed using SAS v9.2 (SAS Institute Inc., Cary, NC).

Statistical methods—group mean response variability

Since some fMRI paradigms may evidence poor within-subject reliability characteristics (e.g., due to habituation or practice effects, such as changes in cognitive strategy), they may be best suited to parallel group designs. In such an experimental design, the relevant comparison is between independent groups of subjects—most simply, two groups and a single session (e.g., treatment and control). Powering such a study relies upon a measure of the expected difference between the two groups under the null hypothesis, i.e., in the absence of an effect. To estimate the distribution of the group-mean differences in the present study, we performed a permutation analysis on the ROI-mean fMRI data from the first session only: the group of N = 25 subjects was arbitrarily split into two sub-groups of N = 12 and N = 13 and both the mean of each group and their difference was calculated. This permutation was repeated 2000 times, generating two distributions. The first was the distribution of the mean sub-group values obtained from the resampling, and indicates the expected spread in group-mean values in a single session. The second was the distribution of the mean difference between the two sub-groups, enabling the calculation of the difference in group means at which statistical significance would be claimed at an alpha = 0.05 level.

Results

Behavioral results

Analyses of the behavioral data revealed that most of the subjects' response data are stable across sessions (Table 2). The only differences, significant only nominally before multiplicity correction at alpha = 5%, occurred in reaction time (RT) during the reward task: Subjects responded faster during session #2 and this difference was mainly driven by the verbal (control) condition. There was also a trend towards significance (p = 0.07) in missing-rate during the n-back task. Here subjects had more misses during session #1 as compared to session #2, mainly driven by the 2-back condition. Overall, reliability of the total RT data was poor for the reward task (ICC = 0.37) and excellent for faces (ICC = 0.84) and the n-back task (ICC = 0.87)—see Table 2 for more details.

fMRI results: group-level consistency across sessions

All three paradigms robustly evoked BOLD signal increases in their respective anatomical target regions (Fig. 2). The group level maps showed substantial overlap in supra-threshold ($p_{FWE}<0.05$) voxels for both the faces and reward tasks within the target ROIs (faces: $R_{OVERLAP} = 0.90$ (left amygdala) and 0.95 (right amygdala); reward: $R_{OVERLAP} = 0.87$ (left VS/NAcc) and 0.97 (right VS/NAcc)). For the n-back task, the extent of supra-threshold voxels was less in the second session but overlap was evident in the right DLPFC ($R_{OVERLAP} = 0.93$ (DLPFC1) and 0.73 (DLPFC2)) and parietal cortex regions ($R_{OVERLAP} = 0.64$ (left parietal) and 0.97 (mid parietal) and 0.81 (right parietal)). Activation effect sizes (ES) for peak voxels within the

Table 2
Behavioral data.

Task	Behavioral measure	Session #1	Session #2	t/p (df = 24)	ICC(2,1) (95%-CI)	ICC(3,1) (95%-CI)
Faces	RT (TOTAL) in ms (±SD)	1091 (205)	1062 (177)	1.33/.20	.83 (.66 .92)	.84 (.66 .92)
	RT (FACES) in ms (±SD)	1150 (244)	1131 (217)	0.73/.47	.85 (.68 .93)	.84 (.68 .93)
	RT (FORMS) in ms (±SD)	1039 (186)	997 (161)	1.58/.13	.69 (.42 .85)	.70 (.43 .86)
	RT (Difference) in ms (±SD)	111 (125)	134 (136)	−0.90/.38	.51 (.15 .75)	.50 (.14 .75)
	Missed (TOTAL) in %	0.42 (1.04)	0.42 (0.85)	0.00/.99	–	–
	Incorrect (TOTAL) in %	1.33 (1.69)	1.08 (1.60)	0.53/.60	–	–
Reward	**RT (TOTAL) in ms (±SD)**[a]	**205 (32)**	**191 (26)**	**2.13/.04**	**.34 (−.02 .64)**	**.37 (−.02 .66)**
	RT (WIN) in ms (±SD)	195 (36)	192 (50)	0.26/.80	.31 (−.09 .63)	.30 (−.09 .62)
	RT (VERBAL) in ms (±SD)	**229 (56)**	**208 (51)**	**2.19/.04**	**.54 (.20 .77)**	**.57 (.24 .79)**
	RT (Difference) in ms (±SD)	−34 (54)	−15 (71)	−1.46/.16	.45 (.09 .71)	.46 (.09 .72)
	Missed in %	–	–	–	–	–
	Incorrect in %	n/a	n/a	n/a	n/a	n/a
	Rewards in € (±SD)	10.88 (2.31)	11.12 (2.09)	−0.53/.60	.48 (.11 .73)	.47 (.11 .73)
n-back	RT (TOTAL) in ms (±SD)	561 (254)	569 (271)	−0.31/.76	.87 (.74 .94)	.87 (.73 .94)
	RT (0-back) in ms (±SD)	591 (212)	578 (230)	0.55/.59	.86 (.71 .94)	.86 (.71 .94)
	RT (2-back) in ms (±SD)	526 (342)	561 (394)	−0.88/.39	.86 (.70 .93)	.85 (.70 .93)
	RT (Difference) in ms (±SD)	−64 (223)	−16 (316)	−1.40/.18	.80 (.60 .90)	.80 (.60 .91)
	Missed (TOTAL) in %	7.05 (6.34)	4.50 (5.33)	1.90/.07	–	–
	Incorrect (TOTAL) in %	8.36 (11.17)	7.56 (13.96)	0.48/.64	–	–
	Missed (0-back) in %	0.21 (0.78)	0 (0)	1.36/.18	–	–
	Incorrect (0-back) in %	0.36 (0.89)	0.14 (0.49)	1.14/.26	–	–
	Missed (2-back) in %	6.83 (6.29)	4.50 (5.33)	1.80/.08	–	–
	Incorrect (2-back) in %	8.00 (11.27)	7.42 (14.01)	0.35/.73	–	–

Note: significant session effects ($p<0.05$, uncorr.) are shown in bold.
[a] RT (TOTAL) for the reward task also includes the loose-condition; n/a = not applicable.

ROIs across the different tasks were generally high (ES = 1.35–2.12 for the faces task, 1.58–1.82 for the reward task, 1.05–2.76 for the n-back task; see Table 3).

Fig. 3 shows the group-level contrast values of session #1 plotted against the contrast-values of session #2 for each voxel in the whole-brain and within the target regions. For all three tasks the second-level activation maps at the whole-brain level are extremely robust (ICCs = 0.88 to 0.98) and this was largely independent of the ICC definition (absolute or relative agreement)—see Table 4.

At the ROI level, relative agreement ICC measure indicates extremely high reliability of all three tasks (mean ICC(3,1) of the whole task battery = 0.87) with somewhat lower values for the faces task (0.72) as compared to the reward (0.94) and the n-back task (0.90). Absolute agreement of ROI contrast-value distribution was lowest for the n-back

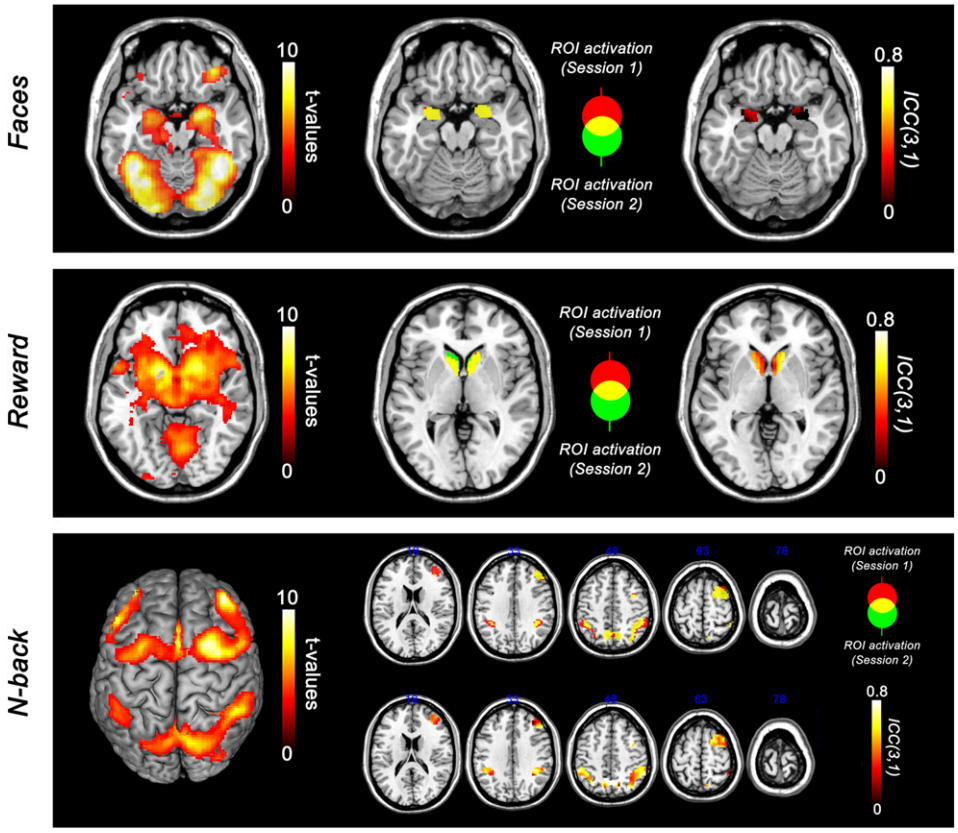


Fig. 2. FMRI activation group level maps from session #1 (left side of each panel). The overlap in supra-threshold ($p_{FWE}<0.05$) voxels within the ROI for the tasks (middle panel) and ICC maps within the particular ROI.

Table 3
fMRI main effects in the regions-of-interest across sessions.

	Session	MNI [a]	k	z-max	t-max	$Mean_{ROI} \pm SD_{ROI}$	$P_{MeanDiff}$	ES_P	ES_R
Faces									
Amygdala L	1	−26 −4 −20	211	5.01	6.75	0.25 (0.23)		1.35	1.11
	2	−22 −4 −22	211	5.54	8.02	0.22 (0.19)	0.30	1.61	1.16
Amygdala R	1	22 −2 −18	248	5.98	9.24	0.24 (0.17)		1.85	1.42
	2	26 −2 −22	248	6.39	10.58	0.23 (0.16)	0.40	2.12	1.44
Reward									
VS/NAcc L	1	−10 6 −8	281	5.50	7.92	1.94 (2.03)		1.59	0.96
	2	−12 8 −8	281	5.93	9.09	2.48 (2.00)	0.08	1.82	1.24
VS/*NAcc R*	1	12 8 −10	277	5.49	7.90	2.41 (2.19)		1.58	1.10
	2	12 8 −8	277	5.88	8.95	3.05 (2.14)	0.05	1.79	1.43
n-back [b]									
DLPFC1 R	1	30 4 60	1052	7.18	13.78	0.65 (0.25)		2.76	2.58
	2	30 4 58	1052	6.07	9.53	0.50 (0.28)	<0.01	1.91	1.76
DLPFC2 R	1	36 44 28	798	6.55	11.15	0.42 (0.22)		2.23	1.91
	2	40 36 28	798	5.18	7.14	0.26 (0.26)	<0.01	1.43	1.01
Parietal mid	1	6 −58 54	717	6.05	9.47	0.81 (0.51)		1.90	1.60
	2	12 −70 54	717	5.90	9.01	0.71 (0.44)	0.14	1.80	1.59
Parietal L	1	−32 50 40	930	6.18	9.86	0.44 (0.28)		1.97	1.57
	2	−32 −50 42	930	4.23	5.23	0.28 (0.34)	<0.05	1.05	0.83
Parietal R	1	38 −46 46	1489	6.54	11.12	0.58 (0.32)		2.23	1.85
	2	42 −46 44	1489	5.13	7.03	0.37 (0.26)	<0.005	1.41	1.40

Note. All reported effects are tested at alpha = 0.05, FWE-corrected for the search volume; k > 10.
MNI = Montreal Neurological Institute coordinates; k = cluster size; ES = effect size (mean beta-parameter divided by its standard deviation); ES_p = effect size for the peak voxel; ES_R = effect size for mean of the total ROI data (i.e. no statistical threshold); $P_{MeanDiff}$ = p-value of the t-test on ROI mean differences.
[a] Only the strongest peak-voxel is listed —see Supplementary Table 5 for additional clusters and/or local maxima within ROI.
[b] ROI mask definitions for the n-back task are empirically derived (see Methods and Supplementary material). Results by anatomical definitions are documented in Supplementary material.

task in the lateral parietal masks (ICC(2,1) = 0.45) and DLPFC2 mask (0.48) while DLPFC1 and medial parietal ROI mask showed excellent reliability (0.75 and 0.96, respectively). For the faces and the reward task, absolute agreement ICC values were all > 0.60.

ROI-mean summary measures showed robust group-level effect sizes in both sessions (ES = 1.11–1.44 for the faces task, 0.96–1.43 for the reward task, 0.83–2.58 for the n-back task; see Table 3). Comparing directly the responses across sessions, stable ROI mean

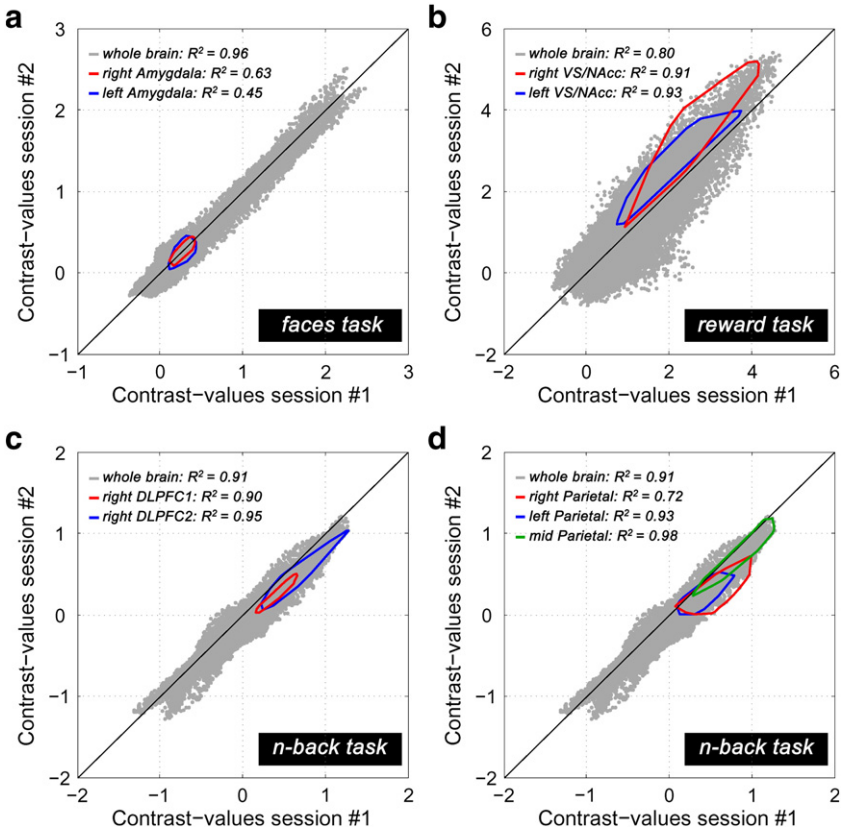


Fig. 3. For all three tasks, the contrast-values of interest of the group results from each voxel in session #1 are plotted against those from session #2 (gray dots). Main diagonals are additionally shown. For the sake of a clear graphical presentation, the scatter plot for each ROI is depicted by its convex hull (outline boundary of the data points). (a) Faces task and amygdala ROIs (k = 211 (left); k = 248 (right)); (b) reward task and VS/NAcc ROIs (k = 281 (left); k = 277 (right)); (c) N-back task and right DLPFC1 (k = 1052), DLPFC2 (k = 798) ROIs and (d) n-back task and parietal ROIs (k = 717 (mid); k = 930 (left); k = 1498 (right)).

Table 4
Group-level map reliability based on voxels (v).

Task	Region	R^2	ICC(2,1)$_v$ (95%-CI) [b]	ICC(3,1)$_v$ (95%-CI) [b]
Faces	wb [a]	.96	.98 (.96 .99)	.98 (.98 .98)
	AMY—L	.45	.62 (.48 .72)	.66 (.57 .73)
	AMY—R	.63	.78 (.72 .83)	.79 (.74 .83)
Reward	wb	.80	.88 (.84 .91)	.89 (.89 .89)
	VS—L	.93	.76 (−.04 .93)	.96 (.95 .97)
	VS—R	.91	.74 (−.06 .92)	.92 (.90 .94)
n-Back	wb	.91	.91 (.59 .96)	.95 (.95 .95)
	DLPFC1—R	.90	.75 (−.06 .93)	.95 (.94 .95)
	DLPFC2—R	.95	.48 (−.01 .82)	.97 (.97 .98)
	Parietal Mid	.98	.96 (.68 .99)	.98 (.98 .99)
	Parietal—L	.60	.45 (−.09 .76)	.77 (.74 .79)
	Parietal—R	.72	.45 (−.07 .78)	.83 (.82 .85)

[a] wb = whole brain.
[b] Note that here BMS refers to between *voxel* mean square.

amplitudes were found for the faces task ($p > 0.30$; paired *t*-test). For the reward task, ROI amplitudes tended to increase slightly across sessions ($p < 0.10$). More significant changes in ROI amplitudes across sessions occurred in the n-back task, where the amplitudes in both right DLPFC masks ($p < 0.01$) as well as in the left ($p < 0.05$) and right ($p < 0.005$) parietal cortex mask were lower in session #2 as compared to session #1. The medial parietal cortex mask amplitudes showed a non-significant decrease ($p > 0.10$) across sessions.

Within-subject reliability across sessions: ROI-level

Poor agreement was found for the faces task regardless of hemisphere and ICC type (consistency or absolute agreement). ICCs did not exceed values of 0.16 for the left amygdala and were zero for the right amygdala. For the reward task, analyses consistently revealed good agreement between sessions regardless of hemisphere and ICC type (ICCs = 0.55–0.62). For the n-back task, we found acceptable reliability for the empirically defined DLPFC1 mask (ICC(3,1) = 0.44), the left parietal ROI (ICC(3,1) = 0.44) and good agreement for the medial parietal ROI (ICC(3,1) = 0.57). The empirically defined DLPFC2 ROI and the right parietal ROI was associated with only poor reliability (ICC = 0.13–0.28) (see Table 5).

Visual inspection of the Bland–Altman plots indicated no systematic dependence of session-to-session differences on the magnitude of the response, in any of the ROIs (see Supplemental Fig. 2). A slight positive bias was evident for both ROIs from the reward task, and a slight negative bias for all ROIs from the n-back task, as described above.

Within-subject reliability across sessions: voxel level

For the faces and reward tasks the ROI-level within-subject reliability was similar to that found at the ROI level. The voxel-level ICCs within the target regions (Fig. 4, Table 5) indicate that, at the voxel scale, the faces task showed poor within-subject reliability (median ICC(3,1) of 0.18 and 0.07 for left and right amygdala respectively) whereas the reward task showed good reliability (median ICC(3,1) = 0.52 and 0.63 for left and right VS/NAcc, respectively). For the n-back task we found high median ICCs for both DLPFC ROIs (ICC(3,1) = 0.58 and 0.44) and all three parietal ROIs (0.59, 0.56 and 0.65 for left, right and mid parietal cortex).

Within-subject reliability versus activation strength

Joint scatter plots revealed task-dependent relationships between group activation strength (in the first session) and within-subject reliability (Fig. 5). For the faces task, the whole brain distribution was skewed toward the top right quadrant, indicating that the most strongly responding voxels also tended to be the most reliable (Fig. 5a). Voxels in the amygdala ROIs were more centrally distributed, reflecting the poor reliability (ICC ~ 0.1) noted above, despite consistent (t ~ 4) albeit weak (contrast ~ 0.25%) group-level activation. The more strongly responding and reliable voxels were primarily localized in the visual cortex, a region also activated by this task although not of primary interest. For the reward task, the whole brain distribution was also skewed toward the top right quadrant; voxels within the ventral striatum target ROIs were also localized in this region of the joint distribution, indicating that the target regions contained among the most strongly responding and reliable voxels engaged in the response to this task (Fig. 5b). For the n-back task, the whole brain distribution was more symmetric by activation strength, being skewed toward the top left and top right quadrants (Fig. 5c). As with the reward task, voxels in the target regions were localized in the top right quadrant, reflecting high activation strength and fair to good reliability.

Inspection of BOLD time courses in voxels with low t-values but high ICCs did not indicate mismodelling but rather stable inter-individual differences in BOLD responses with a range from deactivation to activation leading to non-significant group level activation results (see Supplementary Figs. 3–5).

Effect of gender

For the faces task, we found no substantial differences for male and female subjects at the level of ROI mean amplitude reliability (male: ICC(3,1) = 0.25 and 0.02 (left and right amygdala); female: ICC(3,1) = 0.02 and −0.08). For the reward task we found substantially higher ICCs for females as compared to males (females: ICC(3,1) = 0.83 and 0.82 (left and right VS/NAcc); males: ICC(3,1) = 0.17 and 0.30). For the n-back, we found comparable ICCs for DLPFC1 (males: ICC(3,1) = 0.42; females: 0.47) and DLPFC2 (males: 0.16; females: 0.20). For parietal ROIs, we found slightly higher ICCs for females in mid parietal (females: ICC(3,1) = 0.74; males: 0.42) and right parietal cortex (females: 0.45; males: 0.25).

Include movement parameters as regressors?

So far, all results are based on analyses with movement parameters included in the first level design matrices. Analyses without including movement parameters showed no significant impact on reliability of the reward task. However, for the faces task significantly increased reliability of amygdala activation was observed when movement parameters were not included (ICCs > 0.40). The same was true for the left and right parietal cortex activation during the n-back task (ICCs > 0.40—see Supplementary Table 8 for detailed ICC results). However, statistical analysis of the impact of motion traces on brain activation by means of an overall F-test revealed that task-related head movements were highly reproducible within-subject.

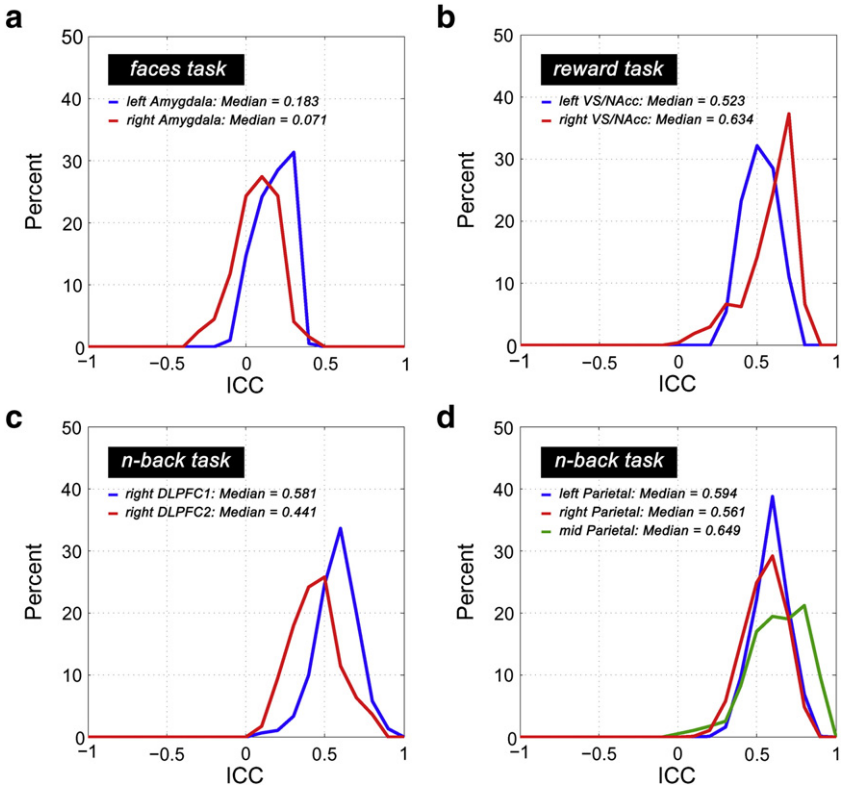


Fig. 4. Distribution of individual voxel ICCs (type: relative agreement, i.e. ICC(3,1)) within each ROI. From top to bottom: (a) faces task with left and right amygdala ICC distribution; (b) reward task with left and right VS/NAcc; (c) n-back with right empirical DLPFC1 and right empirical DLPFC2; (d) n-back with right empirical and left empirical parietal cortex; (e) n-back with mid parietal cortex.

These increased ICC values in the absence of head motion regressors therefore most likely reflect spurious reliability due to stable task related movement.

Stability of group-level response in first session

To assess the stability of group-level responses in a single scanning session, we performed resampling on the session #1 ROI data to assess between subjects reliability. Fig. 6 shows, for each ROI and each task, the distributions of the mean values obtained from resampling the session #1 data into two sub-groups of N = 12 and N = 13 (left), and the distribution of the mean differences between these two resampled subgroups (right). The former indicate the consistency of the group-mean response obtained from a single scanning session per subject. The latter indicate the average difference between means of two independent groups that would be required to detect a significant difference at alpha = 0.05.

Discussion

We have profiled the test–retest reliability of a cognitive-emotive fMRI test battery at both the group and individual subject levels. By investigating the three tasks in the same group of subjects, we were able to attribute differences in task reliability to the tasks themselves with more confidence, rather than to the reliability of the subjects.

We characterized the performance voxel-wise over the whole brain, voxel-wise within the pre-specified ROIs and in terms of the ROI mean summary measures. We also performed an analysis of the robustness of the group-mean response in the first session, to inform the utility of the fMRI paradigms in parallel group designs. These analyses were complemented by an analysis of the behavioral data.

Overall, we found that (1) all three tasks robustly activated their particular target regions; (2) the group-level activation maps were highly stable across sessions for all three tasks; (3) the subject-specific amplitude stability varies considerably for the different

Table 5
Within-subject reliability based on the ROI-mean (m) amplitudes and median (md) of individual voxel ICCs within each ROI (see also Fig. 4).

Task	Region	$ICC(2,1)_m$ (95%-CI)	$ICC(3,1)_m$ (95%-CI)	$ICC(2,1)_{md}$ (5th–95th %ile)	$ICC(3,1)_{md}$ (5th–95th %ile)
Faces	AMY—L	.16 (−.25 .52)	.16 (−.25 .51)	.18 (−.02 .33)	.18 (−.02 .34)
	AMY—R	−.02 (−.43 .38)	−.02 (−.41 .37)	.07 (−.18 .27)	.07 (−.18 .26)
Reward	VS—L	.55 (.22 .77)	.56 (.22 .78)	.52 (.35 .67)	.52 (.35 .67)
	VS—R	.61 (.30 .80)	.62 (.31 .82)	.63 (.26 .76)	.63 (.26 .76)
n-Back	DLPFC1—R	.39 (.03 .67)	.44 (.06 .71)	.57 (.34 .76)	.58 (.36 .77)
	DLPFC2—R	.13 (−.19 .46)	.16 (−.25 .51)	.42 (.19 .71)	.44 (.19 .73)
	Parietal Mid	.57 (.24 .78)	.57 (.23 .78)	.66 (.34 .87)	.65 (.33 .87)
	Parietal—L	.39 (.03 .67)	.44 (.06 .70)	.58 (.39 .74)	.59 (.40 .77)
	Parietal—R	.22 (−.10 .53)	.28 (−.12 .60)	.54 (.31 .73)	.56 (.33 .75)

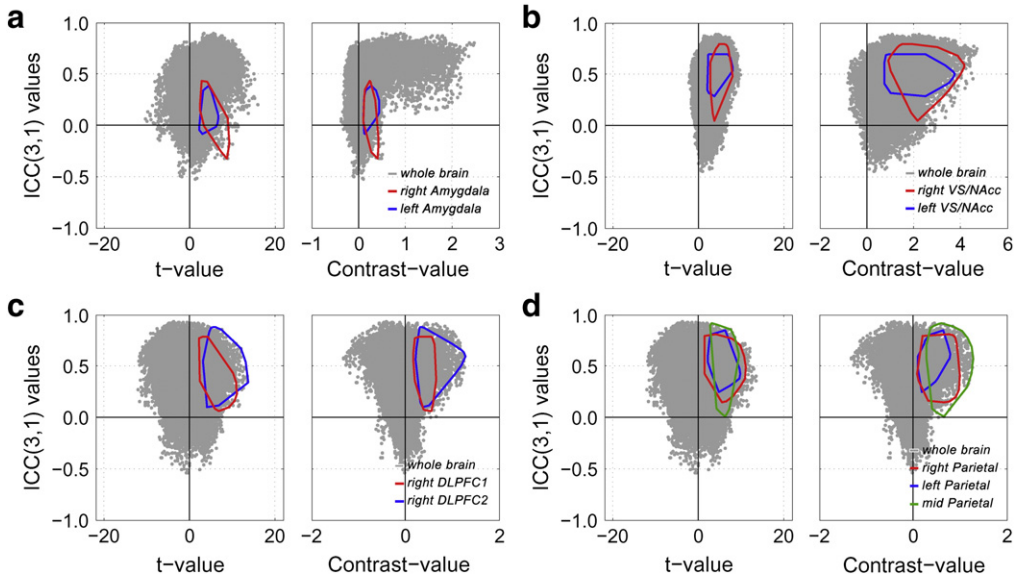


Fig. 5. Voxel-wise correlation of session#1 t-values (left plot of each panel) and contrast-values (right plot of each panel) with the ICC-values for: (a) faces task, (b) reward task, (c) n-back with DLPFC ROI and (d) with parietal cortex ROIs. The scatter plots for each ROI are depicted by their convex hull (outline boundary of the data points).

tasks and ROIs. In the following section we will discuss our results and their implications for future fMRI studies in more detail.

Test–retest reliability: behavioral level

Overall the behavioral data was stable. The only nominally significant difference occurred in the reward task. Here, a decreased RT during the verbal control condition was found while RTs associated with the experimental condition stayed stable. Furthermore, the stability of RT in the reward task was unsatisfactory (ICCs<0.40) and lowest when compared to the other tasks. The first finding is most plausibly a simple training effect and the mean RT stability during the experimental condition may be due to a ceiling effect. The relatively poor stability may be best explained by the fact that the reward task was the only task with an adaptive reaction time window. Therefore, the subject was forced to depart from his/her natural reaction time tendency.

Test–retest reliability: fMRI data

Faces task

No significant group mean ROI amplitude change across sessions occurred during the faces task while the low within-subject amplitude reliability indicates that this is because of heterogeneity in changes across subjects. This might be a consequence of inter-individual differences in emotional processing and emotion regulation strategies and disposition to habituation. Another possible explanation for stable group means but low within-subject reliability is that the faces task is very simple (error rates<1%; presentation duration per trial = 5 s) and therefore the putative off-task time per trial is quite long as derived from the behavioral data (mean RT ~ 1 s). With the current task design

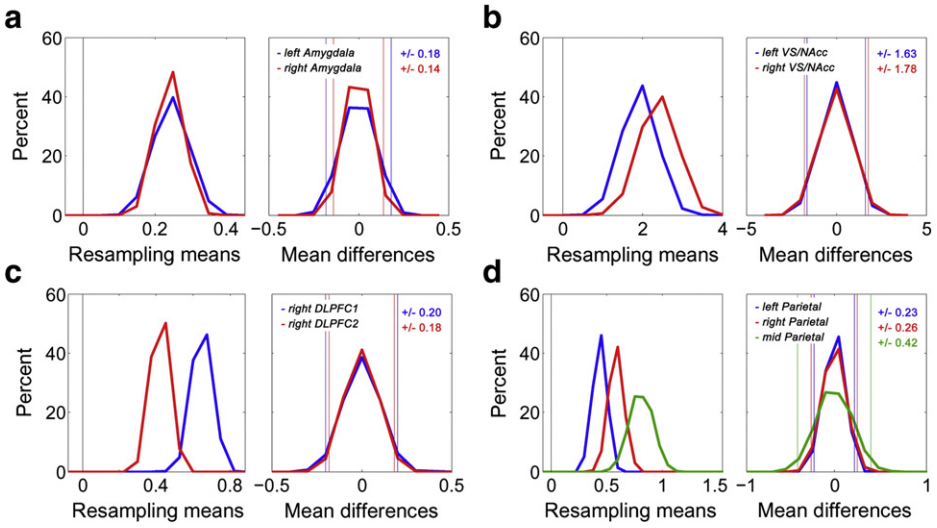


Fig. 6. The two left plots of each panel (a, b, c, d) show the distribution of the ROI mean values (upper left plot: left ROI; lower left plot: right ROI) obtained from the resampling procedure. The two right plots of each panel (a, b, c, d) show the distribution of the mean difference between the random two sub-groups (left and right ROI). Here, the reference lines show the mean difference between means at which statistical significance would be claimed at the 0.05 level.

the mental processes that occurred during the remaining time per trial are not controlled. Large intra-individual differences in ongoing mental activities during off-task time are highly likely and putatively one factor contributing to the low within-subject reliability. If this explanation is valid, the frequently claimed reliability advantage of blocked versus event-related designs (Bennett and Miller, 2010) might not always be valid.

Reward task

At the group-mean level we found a trend toward increased activation in the second session within core structures of the reward system (VS/NAcc). One plausible interpretation might be that subjects tried to win even more money than in session #1, and therefore performed the task with even more rigor. Because we found fair to good ICCs for the reward task, the increase in ROI amplitude across time seem to be relatively consistent over all subjects.

N-back task

When comparing the group-level responses, a notable decrease in activation from session #1 to session #2 was observed in most of the target regions for the n-back task. The main exception was the mid-parietal region which had a stable ROI-mean effect size and also the highest ICC value (0.57). Such a reduction of activation has been associated with training/learning effects (Chein and Schneider, 2005; Ramsey et al., 2004). The behavioral data underpin this interpretation because the lower number of misses in session #2 implies that the subjects were finding the task easier in session #2 as compared to session #1. Despite this group-mean difference between the two sessions, $ICC(3,1)_{ROI}$ values > 0.4 were obtained in three of the five ROIs (DLPFC1, mid-parietal, left-parietal), indicating that the decrease in BOLD response was a reasonably consistent effect across subjects.

Strength of response versus reliability

The joint distributions of t- or contrast values and ICCs (Fig. 5) revealed that the strongest responding voxels were not the most reliable, and vice versa. Indeed, the highest ICC values were observed in voxels with low (~ 0) first session t-values, representing regions of the brain that are reproducibly not engaged by the task. That said, a general association was observed between the strength of response to the task and the within-subject reliability of this response, with the marginal distributions of ICC being skewed toward positive ICC values (i.e., higher absolute values of t being associated with higher ICC values). This is consistent with observations of a similar association but not a one-to-one mapping between strength of response and within-subject reliability at the voxel level with both a working memory and an auditory task (Caceres et al., 2009). In considering this relationship, those authors note that some regions with sub-maximal t-values but high reliability comprise time series that are consistent across sessions but not well modeled by regressors derived from the task paradigm. Caceres et al. (2009) suggested that such regions are involved in response to the task but indirectly or non-linearly to the stimuli. An inspection of our task related BOLD time courses in such voxels with low t-values and high ICCs did not indicate obvious mismodelling of task-induced brain activation but rather stable inter-individual differences in BOLD responses with a range from deactivation to activation and/or task-specific movements (see Supplementary Figs. 3–5).

However, for both the n-back and reward task, voxels in the target regions pre-specified as being of primary interest were located in the upper-right quadrant, toward the extremity of the respective scatter plots, reflecting both strong response to the stimulus as well as reasonable reliability. For the faces task, both within-subject reliability and strength of response were lower for the target regions. However, the magnitude of the BOLD signal change and the t-values (at both voxel and ROI-level) observed here are consistent with those reported using similar facial affect or emotive tasks that have been sufficient for the detection of changes in intervention studies (Del-Ben et al., 2005; Harmer et al., 2006; Murphy et al., 2009). Although it has been consistently shown that emotionally arousing stimuli evoke enhanced activation in the corresponding early sensory cortices (Alpers et al., 2009; Herrmann et al., 2008; Lang et al., 1998; Plichta et al., 2011), the faces paradigm employed here was not balanced in visual field content between the two conditions, leading to additional strong and widespread response in the visual cortex (see Fig. 2). It is these voxels that dominate the upper-right quadrant of Fig. 5a.

Methodological factors

While fair to good within-subject reliability was observed for the reward and n-back tasks, low within-subject reliability estimates were observed in the main target region of the faces task, the amygdala. We tested two effects that potentially could have negatively impacted retest reliability in this task: (a) spatial normalization procedure and (b) amygdala ROI definition.

To test the effect of (a), we re-ran all analyses with indirectly normalized fMRI data (see Methods section) theoretically improving normalization results for small structures. However, although there was a slight increase in activation map reliability for both the left and right amygdala (see Supplementary Table 6), no substantive difference was found for within-subject reliability of the amygdala responses.

To test (b), i.e. alternative ROI definition, we investigated reliability within three subdivisions of the amygdala (Amunts et al., 2005) as well as empirically defined areas within the amygdala according to Johnstone et al. (2005). However, neither procedure led to significantly increased reliability (see Supplementary Table 7 and compare Fig. 5a). Substantially increased reliability of amygdala activation was only observed when movement parameters were not included in the first level design matrices. This, however, most likely reflects spurious reliability because inspection of the motion traces revealed that task related movement was stable in ROIs between sessions.

Because the between-subject analyses indicate robust findings for subjects that performed the task only once, we propose that amygdala habituation might be a plausible reason for the poor within-subject reliability. This is in line with the results of Johnstone et al. (2005) who indicate that habituation of the amygdala due to familiarity of the stimuli might only last for relatively short time periods of 2 weeks but reset with longer time periods. This assumption is also consistent with the long-term reliability of amygdala activation reported by Manuck et al. (2007). Parallel forms (e.g., two comparable sets of emotional faces) of the task stimuli might improve within-subject reliability, but this needs to be demonstrated in future studies. Longer retest-intervals may reset potential habituation effects, but this requirement could prove impractical for use in cross-over studies.

Limitations

The present study focused on reliability of fMRI outputs by applying the widely used ICC index. The results apply to a wide range of healthy volunteer studies (e.g., ph-fMRI, imaging genetics of risk variants) but may not generalize to disease populations (Maiza et al., 2010; Manoach et al., 2001). This is because the ICC is sensitive to the between-subject variance (Bland and Altman, 1990) of a sample, which may be different especially in clinical populations. Examination of the present task battery's reliability in clinical populations is pending.

Reproducibility can also be quantified by “agreement measures” which are independent of the between-subject variance (Bland and Altman, 1999; de Vet et al., 2006). In the present study we also report

95% limits of agreement according to the Bland and Altman method (Bland and Altman, 1986).

Considering gender differences in reliability that were found for the reward task, one should be cautious because of the small sample sizes (10 males and 15 females). However, low reliability for three different reward paradigms has been reported with a sample predominantly consisting of males (Fliessbach et al., 2010). Therefore, future studies should further test the hypothesis that females have more reliable VS/NAcc responses during reward processing than males in this paradigm.

Furthermore, it cannot be excluded that differences in task design affected task reliability. We employed two block-design tasks (faces and n-back) and one event-related design task (reward) which considerably differed with regards to length of task, stimulation density and number of stimulations. Therefore, our intended goal to attribute differences in reliability to the task content (emotive, motivational and cognitive) might be contaminated by effects of task design and duration. Overall, the results argue against this explanation. Although the reward task was nearly twice as long as the other two tasks, the stimulation density (10 trials per condition) was comparably low. This, in turn, makes it unlikely that the longer duration of the reward task was responsible for its high reliability. With regard to design, stimulation density and number of stimulations, the faces and the n-back task are very comparable. However, large differences in reliability occurred between these tasks which are more likely to result from differences in task content itself.

Finally, the ICC(2,1) variant, which tests for absolute agreement, treats session as a random effect. This could potentially be problematic due to the low precision in the estimation of the corresponding variance component. Nevertheless, we found that the ICC(2,1) values were generally very close to the corresponding ICC(3,1) values, implying that this is not a strong confound for the present data.

Recommendations for future studies

All tasks showed excellent group level reliability, making them well suited for between-subject designs, including parallel group pharmacological fMRI (ph-fMRI) and imaging genetics studies. However, the faces task showed poor reliability for within-subject amplitudes. Nevertheless, the group ROI means as well as the permutation results indicate robust amygdala group mean values in the first session. Therefore, a between-subjects (parallel group) design is likely to be preferable when using the faces task in an interventional study, e.g., ph-fMRI.

For the reward task, the VS/NAcc mean amplitudes tended to increase ($p<0.10$) between the first and second sessions with an average (i.e. left and right hemisphere) effect size of $dz = 0.55$. However, the reliability of within-subjects amplitudes in the reward task was fair to good, indicating that this task could be employed in a crossover design if both treatment and session effects are explicitly modeled.

The n-back task also showed excellent group-level consistency, but the mean response amplitudes decreased from session #1 to session #2 and within-subject reliability varied across the ROIs, being fair for empirical DLPFC1 and left and mid parietal cortices. The right parietal and the more anterior DLPFC2 regions exhibited lower reliability. Based on our results, the most sensitive parietal ROI for n-back is the mid parietal region, where reliability was high and intersession amplitude changes were relatively small ($dz = 0.21$). As for the reward task, modeling of the n-back task in a crossover design should include a session effect.

Finally, it is strongly recommended to include movement parameters in the first level design matrices for the faces- and the n-back task.

Summary

All three tasks in the fMRI battery robustly activated their particular target regions and the group-level profiles were all highly stable across sessions. The within-subject reliability varied considerably for the different tasks and ROIs. Both the reward and n-back tasks exhibited fair to good within-subject reliability despite systematic increases and decreases respectively in the BOLD response across the sessions. For these tasks, the order of the sessions does matter and in a within-subject study design systematic retest effects should be taken into account. In contrast, the faces task exhibited stable group-mean response amplitudes across the two sessions but poor within-subject reliability, indicating that this paradigm might be better suited for a between-subject design.

Together, the present study provides task-specific fMRI reliability performance measures that inform the optimal use, powering and design of fMRI studies using comparable tasks.

Acknowledgments

NEWMEDS—the work leading to these results has received funding from the Innovative Medicines Initiative Joint Undertaking (IMI) under grant agreement no. 115008.

We thank Rhiannon Maudsley for her work on the statistical analysis. We thank Georg Gron for generously sharing his expertise in fMRI data analysis. Furthermore, we thank Dagmar Gass for her help with data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2012.01.129.

References

Alpers, G.W., Gerdes, A.B., Lagarie, B., Tabbert, K., Vaitl, D., Stark, R., 2009. Attention and amygdala activity: an fMRI study with spider pictures in spider phobia. J. Neural Transm. 116, 747–757.

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N.J., Habel, U., Schneider, F., Zilles, K., 2005. Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anat. Embryol. (Berl) 210, 343–352.

Barch, D.M., Mathalon, D.H., 2011. Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: psychometric and quality assurance considerations. Biol. Psychiatry 70, 13–18.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? ANYAS 1191, 133–155.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Bland, J.M., Altman, D.G., 1990. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput. Biol. Med. 20, 337–340.

Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 8, 135–160.

Blokland, G.A., McMahon, K.L., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., Wright, M.J., 2011. Heritability of working memory brain activation. J. Neurosci. 31, 10882–10890.

Borsook, D., Becerra, L., Hargreaves, R., 2006. A role for fMRI in optimizing CNS drug development. Nat. Rev. Drug Discov. 5, 411–424.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45, 758–768.

Callicott, J.H., Ramsey, N.F., Tallent, K., Bertolino, A., Knable, M.B., Coppola, R., Goldberg, T., van Gelderen, P., Mattay, V.S., Frank, J.A., Moonen, C.T., Weinberger, D.R., 1998. Functional magnetic resonance imaging brain mapping in psychiatry: methodological issues illustrated in a study of working memory in schizophrenia. Neuropsychopharmacology 18, 186–196.

Chein, J.M., Schneider, W., 2005. Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. Brain Res. Cogn. Brain Res. 25, 607–623.

Del-Ben, C.M., Deakin, J.F., McKie, S., Delvai, N.A., Williams, S.R., Elliott, R., Dolan, M., Anderson, I.M., 2005. The effect of citalopram pretreatment on neuronal responses to neuropsychological tasks in normal volunteers: an fmri study. Neuropsychopharmacology 30, 1724–1734.

de Vet, H.C.W., Terwee, C.B., Knol, D.L., Bouter, L.M., 2006. When to use agreement versus reliability measures. J. Clin. Epidemiol. 59, 1033–1039.

Esslinger, C., Walter, H., Kirsch, P., Erk, S., Schnell, K., Arnold, C., Haddad, L., Mier, D., Opitz von Boberfeld, C., Raab, K., Witt, S.H., Rietschel, M., Cichon, S., Meyer-Lindenberg, A., 2009. Neural mechanisms of a genome-wide supported psychosis variant. Science 324, 605.

Fleiss, J.L. (Ed.), 1986. The Design and Analysis of Clinical Experiments. Wiley, New York.

Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50, 1168–1176.
Forbes, E.E., Brown, S.M., Kimak, M., Ferrell, R.E., Manuck, S.B., Hariri, A.R., 2007. Genetic variation in components of dopamine neurotransmission impacts ventral striatal reactivity associated with impulsivity.
Forbes, E.E., Olino, T.M., Ryan, N.D., Birmaher, B., Axelson, D., Moyles, D.L., Dahl, R.E., 2010. Reward-related brain function as a predictor of treatment response in adolescents with major depressive disorder. Cogn. Affect. Behav. Neurosci. 10, 107–118.
Friedman, L., Glover, G.H., 2006. Report on a multicenter fMRI quality assurance protocol. J. Magn. Reson. Imaging 23, 827–839.
Glahn, D.C., Ragland, J.D., Abramoff, A., Barrett, J., Laird, A.R., Bearden, C.E., Velligan, D.I., 2005. Beyond hypofrontality: a quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. Hum. Brain Mapp. 25, 60–69.
Gountouna, V.-E., Job, D.E., McIntosh, A.M., Moorhead, T.W.J., Lymer, G.K.L., Whalley, H.C., Hall, J., Waiter, G.D., Brennan, D., McGonigle, D.J., Ahearn, T.S., Cavanagh, J., Condon, B., Hadley, D.M., Marshall, I., Murray, A.D., Steele, J.D., Wardlaw, J.M., Lawrie, S.M., 2010. Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49, 552–560.
Hahn, T., Dresler, T., Ehlis, A.-C., Plichta, M.M., Heinzel, S., Polak, T., Lesch, K.-P., Breuer, F., Jakob, P.M., Fallgatter, A.J., 2009. Neural response to reward anticipation is modulated by Gray's impulsivity. Neuroimage 46, 1148–1153.
Hahn, T., Heinzel, S., Dresler, T., Plichta, M.M., Renner, T.J., Markulin, F., Jakob, P.M., Lesch, K.P., Fallgatter, A.J., 2011. Association between reward-related activation in the ventral striatum and trait reward sensitivity is moderated by dopamine transporter genotype. Hum. Brain Mapp. 32, 1557–1565.
Hariri, A.R., Tessitore, A., Mattay, V.S., Fera, F., Weinberger, D.R., 2002. The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage 17, 317–323.
Harmer, C.J., Mackay, C.E., Reid, C.B., Cowen, P.J., Goodwin, G.M., 2006. Antidepressant drug treatment modifies the neural processing of nonconscious threat cues. Biol. Psychiatry 59, 816–820.
Herrmann, M.J., Huter, T., Plichta, M.M., Ehlis, A.C., Alpers, G.W., Muhlberger, A., Fallgatter, A.J., 2008. Enhancement of activity of the primary visual cortex during processing of emotional stimuli as measured with event-related functional near-infrared spectroscopy and event-related potentials. Hum. Brain Mapp. 29, 28–35.
Johnstone, T., Somerville, L.H., Alexander, A.L., Oakes, T.R., Davidson, R.J., Kalin, N.H., Whalen, P.J., 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. Neuroimage 25, 1112–1123.
Kirsch, P., Schienle, A., Stark, R., Sammer, G., Blecker, C., Walter, B., Ott, U., Burkart, J., Vaitl, D., 2003. Anticipation of reward in a nonaversive differential conditioning paradigm and the brain reward system: an event-related fMRI study. Neuroimage 20, 1086–1095.
Kirsch, P., Reuter, M., Mier, D., Lonsdorf, T., Stark, R., Gallhofer, B., Vaitl, D., Hennig, J., 2006. Imaging gene–substance interactions: the effect of the DRD2 TaqIA polymorphism and the dopamine agonist bromocriptine on the brain activation during the anticipation of reward. Neurosci. Lett. 405, 196–201.
Lang, P.J., Bradley, M.M., Fitzsimmons, J.R., Cuthbert, B.N., Scott, J.D., Moulder, B., Nangia, V., 1998. Emotional arousal and activation of the visual cortex: an fMRI analysis. Psychophysiology 35, 199–210.
Lee, J.N., Hsu, E.W., Rashkin, E., Thatcher, J.W., Kreitschitz, S., Gale, P., Healy, L., Marchand, W.R., 2010. Reliability of fMRI motor tasks in structures of the corticostriatal circuitry: implications for future studies and circuit function. Neuroimage 49, 1282–1288.
Liou, M., Su, H.R., Lee, J.D., Cheng, P.E., Huang, C.C., Tsai, C.H., 2003. Functional MR images and scientific inference: reproducibility maps. J. Cogn. Neurosci. 15, 935–945.
Machielsen, W.C., Rombouts, S.A., Barkhof, F., Scheltens, P., Witter, M.P., 2000. FMRI of visual encoding: reproducibility of activation. Hum. Brain Mapp. 9, 156–164.
Maiza, O., Mazoyer, B., Herve, P.Y., Razafimandimby, A., Dollfus, S., Tzourio-Mazoyer, N., Andreassen, O.A., 2010. Impact of cognitive performance on the reproducibility of fMRI activation in schizophrenia. J. Psychiatry Neurosci. 35, 378–389.
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.
Manuck, S.B., Brown, S.M., Forbes, E.E., Hariri, A.R., 2007. Temporal stability of individual differences in amygdala reactivity. Am. J. Psychiatry 164, 1613–1614.
Meyer-Lindenberg, A., 2010. From maps to mechanisms through neuroimaging of schizophrenia. Nature 468, 194–202.
Meyer-Lindenberg, A., Weinberger, D.R., 2006. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nat. Rev. Neurosci. 7, 818–827.
Meyer-Lindenberg, A., Buckholtz, J.W., Kolachana, B., Hariri, A.R., Pezawas, L., Blasi, G., Wabnitz, A., Honea, R., Verchinski, B., Callicott, J.H., Egan, M., Mattay, V., Weinberger, D.R., 2006. Neural mechanisms of genetic risk for impulsivity and violence in humans. Proc. Natl. Acad. Sci. U.S.A. 103, 6269–6274.
Meyer-Lindenberg, A., Straub, R.E., Lipska, B.K., Verchinski, B.A., Goldberg, T., Callicott, J.H., Egan, M.F., Huffaker, S.S., Mattay, V.S., Kolachana, B., Kleinman, J.E., Weinberger, D.R., 2007. Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. J. Clin. Invest. 117, 672–682.
Miki, A., Liu, G.T., Englander, S.A., Raz, J., van Erp, T.G., Modestino, E.J., Liu, C.J., Haselgrove, J.C., 2001. Reproducibility of visual activation during checkerboard stimulation in functional magnetic resonance imaging at 4 Tesla. Jpn. J. Ophthalmol. 45, 151–155.
Muller, R., Buttner, P., 1994. A critical discussion of intraclass correlation coefficients. Stat. Med. 13, 2465–2476.
Murphy, S.E., Norbury, R., O'Sullivan, U., Cowen, P.J., Harmer, C.J., 2009. Effect of a single dose of citalopram on amygdala response to emotional faces. Br. J. Psychiatry 194, 535–540.
Patin, A., Hurlemann, R., 2011. Modulating amygdala responses to emotion: evidence from pharmacological fMRI. Neuropsychologia 49, 706–717.
Pezawas, L., Meyer-Lindenberg, A., Drabant, E.M., Verchinski, B.A., Munoz, K.E., Kolachana, B.S., Egan, M.F., Mattay, V.S., Hariri, A.R., Weinberger, D.R., 2005. 5-HTTLPR polymorphism impacts human cingulate–amygdala interactions: a genetic susceptibility mechanism for depression. Nat. Neurosci. 8, 828–834.
Plichta, M.M., Vasic, N., Wolf, R.C., Lesch, K.-P., Brummer, D., Jacob, C., Fallgatter, A.J., Grön, G., 2009. Neural hyporesponsiveness and hyperresponsiveness during immediate and delayed reward processing in adult attention-deficit/hyperactivity disorder. Biol. Psychiatry 65, 7–14.
Plichta, M.M., Gerdes, A.B., Alpers, G.W., Harnisch, W., Brill, S., Wieser, M.J., Fallgatter, A.J., 2011. Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fNIRS) study. Neuroimage 55, 1200–1207.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532–542.
Ramsey, N.F., Jansma, J.M., Jager, G., Van Raalten, T., Kahn, R.S., 2004. Neurophysiological factors in human information processing capacity. Brain 127, 517–525.
Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Valk, J., Scheltens, P., 1997. Test–retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am. J. Neuroradiol. 18, 1317–1322.
Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16, 105–113.
Rousson, V., Gasser, T., Seifert, B., 2002. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Stat. Med. 21, 3431–3446.
Schacher, M., Haemmerle, B., Woermann, F.G., Okujava, M., Huber, D., Grunwald, T., Kramer, G., Jokeit, H., 2006. Amygdala fMRI lateralizes temporal lobe epilepsy. Neurology 66, 81–87.
Scheres, A., Milham, M.P., Knutson, B., Castellanos, F.X., 2007. Ventral striatal hyporesponsiveness during reward anticipation in attention-deficit/hyperactivity disorder. Biol. Psychiatry 61, 720–724.
Schwarz, A.J., Becerra, L., Upadhyay, J., Anderson, J., Baumgartner, R., Coimbra, A., Evelhoch, J., Hargreaves, R., Robertson, B., Iyengar, S., Tauscher, J., Bleakman, D., Borsook, D., 2011a. A procedural framework for good imaging practice in pharmacological fMRI studies applied to drug development #1: processes and requirements. Drug Discov. Today 16, 583–593.
Schwarz, A.J., Becerra, L., Upadhyay, J., Anderson, J., Baumgartner, R., Coimbra, A., Evelhoch, J., Hargreaves, R., Robertson, B., Iyengar, S., Tauscher, J., Bleakman, D., Borsook, D., 2011b. A procedural framework for good imaging practice in pharmacological fMRI studies applied to drug development #2: protocol optimization and best practices. Drug Discov. Today 16 (15/16), 671–682.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. PsyB 86, 420–428 (URLJ) http://www.apa.org/journals/bul.html.
Specht, K., Willmes, K., Shah, N.J., Jancke, L., 2003. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 17, 463–471.
Stark, R., Schienle, A., Walter, B., Kirsch, P., Blecker, C., Ott, U., Schafer, A., Sammer, G., Zimmermann, M., Vaitl, D., 2004. Hemodynamic effects of negative emotional pictures — a test–retest analysis. Neuropsychobiology 50, 108–118.
Tan, H.Y., Chen, Q., Sust, S., Buckholtz, J.W., Meyers, J.D., Egan, M.F., Mattay, V.S., Meyer-Lindenberg, A., Weinberger, D.R., Callicott, J.H., 2007. Epistasis between catechol-O-methyltransferase and type II metabotropic glutamate receptor 3 genes on working memory brain function. Proc. Natl. Acad. Sci. U.S.A. 104, 12536–12541.
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.
Wagner, K., Frings, L., Quiske, A., Unterrainer, J., Schwarzwald, R., Spreer, J., Halsband, U., Schulze Bonhage, A., 2005. The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. Neuroimage 28, 122–131.
Wei, X., Yoo, S.S., Dickey, C.C., Zou, K.H., Guttmann, C.R., Panych, L.P., 2004. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. Neuroimage 21, 1000–1008.
Wise, R.G., Preston, C., 2010. What is the value of human FMRI in CNS drug development? Drug Discov. Today 15, 973–980.
Wise, R.G., Tracey, I., 2006. The role of fMRI in drug discovery. J. Magn. Reson. Imaging 23, 862–876.
Yetkin, F.Z., McAuliffe, T.L., Cox, R., Haughton, V.M., 1996. Test–retest precision of functional MR in sensory and motor task activation. AJNR Am. J. Neuroradiol. 17, 95–98.