Published in final edited form as:
Neuroimage. 2006 February 1; 29(3): 1000–1006.

Long-term test–retest reliability of functional MRI in a classification learning task

Adam R. Aron, **Mark A. Gluck**, and **Russell A. Poldrack***
Department of Psychology, Franz Hall Box 951563, University of California, Los Angeles, CA 90095, USA

Abstract

Functional MRI is widely used for imaging the neural correlates of psychological processes and how these brain processes change with learning, development and neuropsychiatric disorder. In order to interpret changes in imaging signals over time, for example, in patient studies, the long-term reliability of fMRI must first be established. Here, eight healthy adult subjects were scanned on two sessions, 1 year apart, while performing a classification learning task known to activate frontostriatal circuitry. We show that behavioral performance and frontostriatal activation were highly concordant at a group level at both time-points. Furthermore, intra-class correlation coefficients (ICCs), which index the degree of correlation between subjects at different time-points, were high for behavior and for functional activation. ICC was significantly higher within the network recruited by learning than outside that network. We conclude that fMRI can have high long-term test–retest reliability, making it suitable as a biomarker for brain development and neurodegeneration.

Keywords

Probabilistic classification learning; Basal ganglia; Neurodegenerative disease; Longitudinal study; Intra-class correlation

Introduction

Functional magnetic resonance imaging (fMRI) has become the method of choice for non-invasive imaging of human cognitive functions. Recent work has strongly linked fMRI signals to synaptic activity and neuronal firing (reviewed in Logothetis, 2003), and these data are confirmed by convergent effects of stimulus manipulations (e.g., contrast of visual stimuli) across both fMRI and neurophysiological techniques (Rees et al., 2000). However, both the *validity* and *reliability* of fMRI for measuring signals relevant to higher cognitive function continue to be questioned (e.g., Uttal, 2001). Regarding the reliability of fMRI signals, we note that meta-analyses generally do find impressive *concordance* across studies, though there is often substantial variability as well (e.g., Buchsbaum et al., 2005; Derrfuss et al., 2005; Duncan and Owen, 2000; Ridderinkhof et al., 2004; Wager et al., 2004; Wager and Smith, 2003). Another aspect of reliability, which we investigate here, regards the *reproducibility* of fMRI signals over different scanning sessions.

A number of prior studies have examined the reproducibility of fMRI signals in experiments of visual stimulation (Rombouts et al., 1997), fear and disgust (Stark et al., 2004), auditory odd-ball processing (Kiehl and Liddle, 2003), working memory (Manoach et al., 2001; Wei et al., 2004) and sensorimotor control (Loubinoux et al., 2001; Yoo et al., 2005). For five of these

* Corresponding author. *E-mail address:* poldrack@ucla.edu (R.A. Poldrack)..

studies, the test–retest interval was only on the short-term (i.e. for an inter-session interval of at most a few weeks). Moreover, some of these studies used an approach where they compared either group activation maps or single-subject activation maps at different time-points. Comparing activation maps in this way is not ideal for establishing test–retest reliability of fMRI signals (McGonigle et al., 2000; Poline et al., 1996). The problem is that thresholding of images can exaggerate very small differences between maps: the signal level could be highly reliable, yet small differences in the signal or noise could result in substantial differences in thresholded maps due to the nonlinearity of the thresholding operation. A more promising approach is to extract signal change for each subject at each time-point and compute intra-class correlation coefficients (ICCs) to assess reliability (cf. Manoach et al., 2001). The question then revolves around how to choose the regions of interest from which to extract signal change. This could be achieved either by using a priori defined regions, or, as we do here, by extracting ICC values from the network that is activated at a group level for either session 1 or session 2 (inclusively).

Two other studies examined test–retest reliability over the longer term, studying subjects at 9 time-points with at least 3 weeks between sessions (Wei et al., 2004; Yoo et al., 2005). Wei et al. examined a working memory paradigm and showed that session maps were consistent across time. However, as they did not model subject as a random effect, the results are not generalizable outside of that particular sample. Yoo et al. examined a finger-tapping paradigm, using group activation maps to localize three ROIs in the motor system. Again, the authors used an approach in which mean activation for each subject within each ROI was then computed (moreover in native, not standard, space). There was substantial variability in volume and spatial distribution of activation across sessions, suggesting that, for this task and/or method, test–retest reliability of fMRI signals was not high.

In summary, no study has yet demonstrated a high correlation in functional activation across subjects between two or more sessions over the long-term. This is a serious methodological lacuna, as such reliability must be established before fMRI can be effectively deployed to study long-term learning, development, neurodegeneration or treatment (Casey et al., 2005; Paulsen et al., 2004). For example, in Huntington's disease research, the time is nearing when treatments from mouse models may be translated to human clinical trials (Beal and Ferrante, 2004). As such treatments may be designed to protect neurons before they degenerate, fMRI, rather than PET or structural MRI, may be the method of choice for judging the functional integrity of brain networks in response to a cognitive task. Yet, the interpretation of longitudinal changes in fMRI signals in such studies first requires that the measures be shown to be reliable over time in healthy volunteers.

The present study aimed to establish test–retest reliability for functional MRI using a complex cognitive task that engages broad networks in the brain, rather than discrete foci. As such, this task could be useful in assessing longitudinal change in neurodegenerative conditions characterized by changes to such networks as the frontostriatal system. Another important characteristic of a candidate task is that it is shown to exhibit minimal *practice effects*, wherein behavioral scores improve over time as subjects become more practiced at the task they are performing. As practice effects are also associated with changes in observed fMRI signal (reviewed in Kelly and Garavan, 2005), it is clearly important to choose a task with minimal practice effects in order to assess test–retest reliability of fMRI signals (cf. Manoach et al., 2001; McGonigle et al., 2000).

We employed a probabilistic classification learning (PCL) task which met these desiderata. PCL is a difficult problem of classification which requires subjects to learn on the basis of trial-by-trial feedback (Fig. 1). We studied eight subjects in two fMRI scanning sessions separated by just over 1 year. The nature of the task was identical for the two sessions, but the material

to be learned changed in each session. Although it is possible that subjects could develop skill or strategy in how they go about learning a particular classification, pilot data suggested this would not affect the accuracy of their classifications for *new* materials. Hence, we expected that practice effects between the two versions of the task would be minimal. Furthermore, we had already established that, when PCL trials are contrasted with baseline (non-learning) trials, a network of midbrain, striatal and frontal regions, consistent with the mesencephalic dopamine system, is robustly activated (Aron et al., 2004;Poldrack et al., 2001). In the current study, we compared the level of activation across sessions within this frontostriatal network and computed intra-class correlations to quantify the level of reliability. The results demonstrate that fMRI signals in the frontostriatal system are highly reliable over the two sessions.

Methods

Subjects

Eight right-handed healthy English-speaking subjects participated twice each (3 males/5 females; age range 21–26 years; mean age 23.25 ± 1.83; mean interval between scans 13.5 ± 0.93 months). All subjects were carefully screened to make sure they had no history of neurological or psychiatric disorder. All subjects gave informed consent according to a UCLA Institutional Review Board protocol.

Behavioral task

Subjects performed a classification learning task, which has been extensively studied previously (e.g., Aron et al., 2004; Beninger et al., 2003; Keri et al., 2002; Knowlton et al., 1994, 1996; Moody et al., 2004; Poldrack et al., 2001; Shohamy et al., 2004). On each trial, one to three (out of 4 potential) cards were presented: giving 14 potential different combinations (we used just 13 of these). The location of the cards was random. Each of the combinations constituted a 'stimulus,' and the subject had to indicate whether the outcome would be sun (left button press) or rain (right button press). The probability with which each stimulus was associated with rain is shown in Table 1. Frequencies were chosen in such a way that the cue–outcome associations (i.e. the associations between each *particular* card and the rain outcome) were 0.18, 0.37, 0.59 and 0.82; these probabilities are similar but slightly more deterministic than previous studies (e.g., Knowlton et al., 1994). Therefore, both individual cue–outcome associations as well as configuration–outcome associations were generally probabilistic.

For each experimental session, there were 100 PCL trials, randomized for each participant, and these were presented in two scanning runs of 50 PCL trials each. In addition, each scanning run contained 30 baseline trials for fMRI analysis purposes (i.e. to control for visual stimulation, response and feedback). In each scanning run of 80 trials total, there were 10 cycles consisting of 5 consecutive PCL trials followed by 3 consecutive trials of a baseline task (Fig. 1a). Stimulus presentation lasted for 3 s, within which time the subject responded with a left button press for sun or a right button press for rain. As soon as the subject responded, feedback (the word "rain" or "sunshine") was presented along with the stimulus (the default was that feedback presentation lasted for 1 s) (Fig. 1b). There was a 0.5-s second interstimulus interval. Baseline trials consisted of a standard pattern at all three card positions for 3 s, along with the instruction "press" (Fig. 1c). The subject was instructed to always press the right button on baseline trials. As soon as the button was pressed, the word "press" disappeared.

Procedure

In each session, subjects were briefly practiced on one cycle (5 PCL trials, randomly chosen, and 3 baseline trials) outside the scanner to familiarize them with task requirements. It was emphasized that the left key should be pressed with the left index finger for a prediction of 'sunshine' and the right key with the right index finger for a prediction of 'rain.' It was

explained to the subject that s/he would be guessing at first, but should respond on every trial, that location of the cards was not important and that cycles would be presented of 5 PCL trials followed by 3 baseline trials. Once in the scanner, subjects performed two scanning runs (80 trials each, 4.5 s per trial, 6-min duration) with a short break between scans. Subjects used left and right index fingers to press left and right buttons on the MR-compatible button box. The only difference in procedure between the two scanning sessions was that the color and shapes making up the stimuli changed in order to prevent transference of learning to the second session (Fig. 1d). For each session, the assignment of the four cards to each of the four cues was pseudo-randomized across subjects.

Behavioral analyses

Accuracy was estimated with a 'maximizing metric' by assessing whether the subject's response was correct with respect to p(rain) for each of the 13 stimulus types (cf. Knowlton et al., 1994). A response for a particular PCL trial counted as correct if p(rain) > 0.5 and the subject pressed the key for rain or if p(rain) < 0.5 and the subject pressed the key for sunshine [p(rain) was computed over all 100 trials]. If p(rain) equaled 0.5 (for one stimulus type), the trial was excluded from behavioral analysis. Percent correct scores were computed for each subject for each block/scan of each session and entered into ANOVA (2 sessions × 2 blocks) with subject as a random factor. Additionally, reliability of behavioral scores was computed (for the scan blocks 1 and 2) at the two time-points using the intra-class correlation coefficient (ICC; see Reliability analyses section below). (Note: behavioral data and scan data were missing from one block for one subject on the second session, so this subject's data were not entered into ANOVA but were used for computing ICC for scan 1.)

MRI data acquisition

A 3 T Siemens Allegra MRI scanner was used to acquire 180 functional T2*-weighted echoplanar images (EPI) (4 mm slice thickness, 33 slices, TR = 2 s, TE = 30 ms, flip angle = 90°, matrix 64 × 64, field of view 200). Stimulus presentation and timing of all stimuli and response events were achieved using MATLAB (http://www.mathworks.com) and the Psychtoolbox (http://www.psychtoolbox.org). Additionally, a matched-band-width High-Resolution scan (same slice prescription as EPI) and MPRAGE were acquired for each subject for registration purposes. The MPRAGE had parameters: TR = 2.3, TE = 2.1, FOV = 256, matrix = 192 × 192, saggital plane, slice thickness = 1 mm, 160 slices.

Imaging analysis

Identical methods were used for analysis of functional MRI data for the two scanning sessions. Initial analysis was carried out using tools from the FMRIB software library (http://www.fmrib.ox.ac.uk/fsl). The first two volumes were discarded to allow for T1 equilibrium effects. The remaining images were then realigned to compensate for small head movements (Jenkinson et al., 2002) and were spatially smoothed using a 5-mm full-width half-maximum (FWHM) Gaussian kernel. Translational movement parameters never exceeded 0.5 of a voxel in any direction for any subject or session. The data were filtered in the temporal domain using a nonlinear high-pass filter with a 66-s cut-off. A three-step registration procedure was used whereby EPI images were first registered to the matched-bandwidth High-Resolution scan, then to the MPRAGE structural image and finally into standard (MNI) space, using affine transformations (Jenkinson and Smith, 2001).

For each scan, PCL trials alone were modeled after convolution with a canonical hemodynamic response function. A nuisance regressor was added, which consisted of trials on which no response was made (usually fewer than 5% trials). Temporal derivatives were included as covariates of no interest to improve statistical sensitivity. This procedure produced, for each

subject, each scan and each session, a contrast image of PCL trials vs. implicit (unmodeled) baseline.

For each subject, the two contrast images for each session were averaged, giving 8 such images for each session (for the one subject who only had one scan from the second session, this scan alone was used). A random effects statistical analysis was carried out on the contrast images separately for each session. Group images were thresholded using cluster detection statistics, with a height threshold of $z > 2.3$ and a cluster probability of $P < 0.01$, corrected for multiple comparisons (using Gaussian Random Field Theory).

Reliability analyses

Custom MATLAB code was written to compute ICC on a voxel-by-voxel fashion for the 8 contrast images at the two time-points. ICC was computed as:

$$\mathrm{ICC} = (\mathrm{MSEbetwsubs} - \mathrm{MSEwithinsubs}) / (\mathrm{MSEbetwsubs} + \mathrm{MSEwithinsubs})$$

Where MSEbetwsubs and MSEwithinsubs are the mean square errors for between-subjects and within-subjects variance respectively (where these values are taken from a repeated measures ANOVA with 8 subjects and two session variables, i.e. sessions 1 and 2). ICC represents the ratio of between-subject variance to total variance and is the appropriate metric for assessing within-subject reliability, rather than Pearson's *R*, because the observations are not independent (Shrout and Fleis, 1979). Therefore, ICC values will be particularly high when within-subject (i.e. within-subject between-session) variance is low *and* between-subject variance is high. The resulting 3D voxel map of ICC values (>0.5) was then masked (by multiplication) with a binary image representing the PCL network activated for either session 1 *or* session 2: that is, we created a binary PCL mask using the group maps from session 1 and session 2, voxel thresholded at $z > 2.3$ with a cluster probability of $P < 0.01$, corrected for multiple comparisons. ICC values are therefore only displayed within brain regions activated by PCL in session 1 *or* session 2. A final analysis used a Chi-square test to assess whether there were significant differences between the distribution of ICC values within the PCL network compared to the distribution in brain regions outside that network (exPCL).

Results

Behavior

There was a main effect of learning, so that, within sessions, accuracy was significantly greater for block 2 than for block 1, $F(1,6) = 116.4$, $P < 0.0001$ (Fig. 2a) (Note: data missing for one block for one subject reduce *df* from 7 to 6, see Methods.) However, there was no significant difference in accuracy between sessions, $F(1,6) = 1.1$, n.s. [session 1: 74.9%, session 2: 76.4%], and the interaction between block and session was not significant, $F(1,6) = 1.1$, n.s. Across subjects, average accuracy for session 1 and session 2 was highly correlated; $df = 6$, ICC = 0.8514, $P = 0.0037$, and this was also the case for scan 1 in session 1 vs. scan 1 in session 2 ($df = 7$, ICC = 0.64, $P < 0.05$), and scan 2 in session 1 vs. scan 2 in session 2 ($df = 6$, ICC = 0.77, $P < 0.05$) (Figs. 2b, c). The stability of behavior in scan 1, across sessions, and scan 2, across sessions, was confirmed by non-significant sign tests (both $P > 0.7$).

Group activation maps

For the contrast of PCL trials minus baseline trials, both session 1 and session 2 produced significant activation of frontostriatal circuitry (caudate, putamen, globus pallidus, thalamus, orbital, lateral and medial frontal cortex) as well as midbrain, consistent with our prior results using somewhat different behavioral and analysis procedures (Aron et al., 2004; Poldrack et

al., 2001) (Fig. 3). For a direct comparison of this contrast between sessions, there was significantly more activation for session 2 than session 1 in right dorsal anterior PFC (MNI: 28 50 32 [*x y z*], $t = 12.7$) and left dorsal PFC (MNI: −28 34 22 [*x y z*], $t = 7.34$). There were no regions for which activation was significantly greater for session 1 than session 2.

Intra-class correlations for fMRI

ICC values within the network significantly activated by PCL for session 1 *or* session 2 were high, often exceeding 0.8 (Fig. 4a). This is illustrated for key ROIs: across subjects, mean effect size for the comparison of the classification task with the baseline task in midbrain, striatal and frontal ROIs is highly correlated for session 1 compared to session 2 (Figs. 4b, c). ICC values were significantly higher for voxels within the PCL network compared to voxels outside the network [Chi-square (9 *df*, for 10 intervals per distribution) = 781, $P < 0.0001$] (Fig. 5).

Discussion

The results show that fMRI can have high test–retest reliability over the long-term. In particular, activation within the frontostriatal network known to underlie classification learning was highly consistent across the two sessions, as assessed by intra-class correlations. This has direct implications for assessing longitudinal change as a function of development, neuropsychiatric disorders or treatment.

Subjects were studied on two occasions, on the same scanner, separated by just over 1 year. Preprocessing and analysis of imaging data were identical between sessions, and subject head movement was always minimal. The only differences between sessions pertained to the color and shape of the features to be classified in the PCL task and the order of trials. Learning in both sessions robustly activated the frontostriatal network, as we have seen in prior studies using somewhat different behavioral and analysis procedures (Aron et al., 2004; Poldrack et al., 2001). A direct comparison between sessions showed increased activation at two frontal foci for session 2 vs. 1, but not for session 1 vs. 2. These foci could represent regions of plasticity related to task strategy, rather than learning of the material, as the foci were not consistent with the network activated by learning and learning performance across sessions was highly correlated among subjects (and mean performance between sessions equivalent). Therefore, there were no significant differences in activation within the learning network between sessions, and the increase of activation at frontal foci outside this network for session 2 probably represents neural plasticity related to task strategy rather than a change in classification learning itself.

We further examined ICC values within the network associated with PCL. ICC values were very high, as reflected in the scatterplots of mean signal, at key frontal, midbrain and striatal foci, confirming the reliability of test–retest activation at these foci. Furthermore, it was unlikely that this result arose merely because subjects who activated highly in session 1 also activated highly in session 2 (e.g., due to global changes in SNR) as ICC values within the PCL network were significantly higher than for ICC values outside the PCL network. Therefore, the high test–retest reliability for functional activation was fairly specific to the network known to underlie PCL performance from neuroimaging and neuropsychology (Aron et al., 2004; Beninger et al., 2003; Keri et al., 2002; Knowlton et al., 1996; Knowlton et al., 1994; Moody et al., 2004; Poldrack et al., 2001; Shohamy et al., 2004). One way to apply this method in a longitudinal study of neurodegenerative disease or treatment is to extract mean signal from ROIs within this network and to assess statistically whether differences between test and retest activation interact with group (e.g., patients vs. controls, or drug vs. placebo). Two important caveats in any patient group, however, would be that the patients did show significant learning of the task and that they had roughly similar variance in their fMRI data compared to controls.

A limitation of the study is that we have only established test–retest reliability of fMRI signals for one task. An open question is whether this study could be repeated for a range of cognitive paradigms such as those requiring motor learning or executive control, which are well known to activate frontostriatal and other networks in the brain, and may also serve as reliable biomarkers. Our results here, combined with a consideration of the studies that have examined test–retest reliability of fMRI signals across shorter time-spans, as well as the literature on practice effects in fMRI, strongly suggest that candidate cognitive tasks should first be shown to have minimal behavioral practice effects across time, before fMRI reliability is evaluated. Furthermore, our results strongly motivate an approach to assessing test–retest reliability based on computing signal change and comparing this across subjects using ICC, as opposed to the use of thresholded maps. In this study, we computed signal change within the PCL network significantly activated by session 1 or session 2. Future studies on the same scanner, studying subjects of the same mean age and employing the same task and analysis method, could use the PCL network activated here as an a priori region of interest.

This study fills a methodological lacuna by showing high behavioral and functional MRI test–retest reliability for the PCL task within a frontostriatal system at a 1-year interval. As clear predictions can be made regarding longitudinal change in fMRI signals for this task in the frontostriatal system in patients with Huntington's disease, Parkinson's disease, obsessive-compulsive disorder and schizophrenia (e.g., Beninger et al., 2003; Keri et al., 2002; Knowlton et al., 1994, 1996; Moody et al., 2004; Rauch et al., 1997; Shohamy et al., 2004), we have supplied a method that is readily applicable to assessing neurodegeneration and neuro-protection in these groups in comparison with appropriate age-matched control subjects.

Acknowledgements

Whitehall Foundation and NSF grant BCS-0223843 to R.A.P. The authors thank Allan J. Tobin and Robert Bilder for helpful discussion and encouragement, Sabrina Tom for scanning and Catherine Myers and Daphna Shohamy for help with task design.

References

Aron AR, Shohamy D, Clark J, Myers C, Gluck MA, Poldrack RA. Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. J Neurophysiol 2004;10:10.

Beal MF, Ferrante RJ. Experimental therapeutics in transgenic mouse models of Huntington's disease. Nat Rev, Neurosci 2004;5:373–384. [PubMed: 15100720]

Beninger RJ, Wasserman J, Zanibbi K, Charbonneau D, Mangels J, Beninger BV. Typical and atypical antipsychotic medications differentially affect two nondeclarative memory tasks in schizophrenic patients: a double dissociation. Schizophr Res 2003;61:281–292. [PubMed: 12729880]

Buchsbaum BR, Greer S, Chang WL, Berman KF. Meta-analysis of neuroimaging studies of the Wisconsin Card-Sorting task and component processes. Hum Brain Mapp 2005;25:35–45.

Casey BJ, Galvan A, Hare TA. Changes in cerebral functional organization during cognitive development. Curr Opin Neurobiol 2005;15:239–244. [PubMed: 15831409]

Derrfuss J, Brass M, Neumann J, von Cramon DY. Involvement of the inferior frontal junction in cognitive control: meta-analyses of switching and Stroop studies. Hum Brain Mapp 2005;25:22–34. [PubMed: 15846824]

Duncan J, Owen AM. Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends Neurosci 2000;23:475–483. [PubMed: 11006464]

Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5:143–156. [PubMed: 11516708]

Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 2002;17:825–841. [PubMed: 12377157]

Kelly AM, Garavan H. Human functional neuroimaging of brain changes associated with practice. Cereb Cortex 2005;15 (8):1089–1102. [PubMed: 15616134]

Keri S, Szlobodnyik C, Benedek G, Janka Z, Gadoros J. Probabilistic classification learning in Tourette syndrome. Neuropsychologia 2002;40:1356–1362. [PubMed: 11931939]

Kiehl KA, Liddle PF. Reproducibility of the hemodynamic response to auditory oddball stimuli: a six-week test–retest study. Hum Brain Mapp 2003;18:42–52. [PubMed: 12454911]

Knowlton BJ, Squire LR, Gluck MA. Probabilistic classification learning in amnesia. Learn Mem 1994;1:106–120. [PubMed: 10467589]

Knowlton BJ, Mangels JA, Squire LR. A neostriatal habit learning system in humans. Science 1996;273:1399–1402. [PubMed: 8703077]

Logothetis N. The underpinnings of the BOLD functional magnetic resonance imaging signal. J Neurosci 2003;23:3963–3971. [PubMed: 12764080]

Loubinoux I, Carel C, Alary F, Boulanouar K, Viallard G, Manelfe C, et al. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. J Cereb Blood Flow Metab 2001;21:592–607. [PubMed: 11333370]

Manoach DS, Halpern EF, Kramer TS, Chang Y, Goff DC, Rauch SL, et al. Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am J Psychiatry 2001;158:955–958. [PubMed: 11384907]

McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RS, Holmes AP. Variability in fMRI: an examination of intersession differences. NeuroImage 2000;11:708–734. [PubMed: 10860798]

Moody TD, Bookheimer SY, Vanek Z, Knowlton BJ. An implicit learning task activates medial temporal lobe in patients with Parkinson's disease. Behav Neurosci 2004;118:438–442. [PubMed: 15113271]

Paulsen JS, Zimbelman JL, Hinton SC, Langbehn DR, Leveroni CL, Benjamin ML, et al. fMRI biomarker of early neuronal dysfunction in presymptomatic Huntington's disease. AJNR Am J Neuroradiol 2004;25:1715–1721. [PubMed: 15569736]

Poldrack RA, Clark J, Pare-Blagoev EJ, Shohamy D, Moyano JC, Myers C, et al. Interactive memory systems in the human brain. Nature 2001;414:546–550. [PubMed: 11734855]

Poline JB, Vandenberghe R, Holmes AP, Friston KJ, Frackowiak RS. Reproducibility of PET activation studies: lessons from a multi-center European experiment. EU concerted action on functional imaging NeuroImage 1996;4:34–54.

Rauch SL, Savage CR, Alpert NM, Dougherty D, Kendrick A, Curran T, et al. Probing striatal function in obsessive-compulsive disorder: a PET study of implicit sequence learning. J Neuropsychiatry Clin Neurosci 1997;9:568–573. [PubMed: 9447498]

Rees G, Friston K, Koch C. A direct quantitative relationship between the functional properties of human and macaque V5. Nat Neurosci 2000;3:716–723. [PubMed: 10862705]

Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S. The role of the medial frontal cortex in cognitive control. Science 2004;306:443–447. [PubMed: 15486290]

Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Valk J, Scheltens P. Test–retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am J Neuroradiol 1997;18:1317–1322. [PubMed: 9282862]

Seger CA, Cincotta C. The roles of the caudate nucleus in human classification learning. J Neurosci 2005;16:2941–2951. [PubMed: 15772354]

Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA, Poldrack RA. Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. Brain 2004;127 (pt 4):851–859. [PubMed: 15013954]

Shrout PE, Fleis J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;2:420–428.

Stark R, Schienle A, Walter B, Kirsch P, Blecker C, Ott U. Hemodynamic effects of negative emotional pictures—a test–retest analysis. Neuropsychobiology 2004;50:108–118. [PubMed: 15179028]

Uttal, WR. The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain. The MIT Press: Cambridge, MA.; 2001.

Wager TD, Smith EE. Neuroimaging studies of working memory: a meta-analysis. Cogn Affect Behav Neurosci 2003;3:255–274. [PubMed: 15040547]

Wager TD, Jonides J, Reading S. Neuroimaging studies of shifting attention: a meta-analysis. NeuroImage 2004;22:1679–1693. [PubMed: 15275924]

Wei X, Yoo SS, Dickey CC, Zou KH, Guttmann CR, Panych LP. Functional MRI of auditory verbal working memory: long-term reproducibility analysis. NeuroImage 2004;21:1000–1008. [PubMed: 15006667]

Yoo SS, Wei X, Dickey CC, Guttmann CR, Panych LP. Long-term reproducibility analysis of fMRI using hand motor task. Int J Neurosci 2005;115:55–77. [PubMed: 15768852]

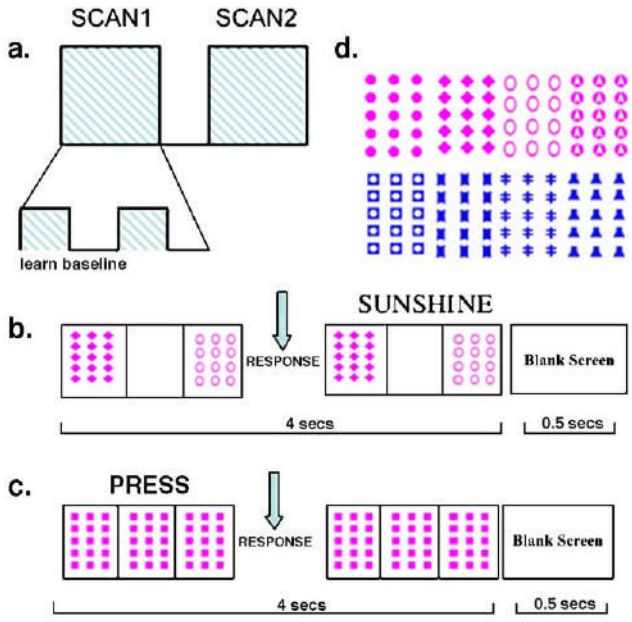

Fig. 1.

Scanning design for probabilistic classification learning (PCL) and baseline trials. (a) On each occasion (session), subjects performed 2 scans, each consisting of 10 cycles of 5 PCL trials and 3 baseline trials (80 trials total per scan). (b) On each weather prediction trial, a stimulus was presented, comprising 1 to 3 cards, at randomized locations, for up to 4 s. Within that time, the subject responded with left button press (sun) or right button press (rain). Feedback ("sunshine" or "rain") was presented after button press for the remainder of the 4-s window. Intertrial interval was 0.5 s. (c) Baseline trials controlled for visual stimulation, button press and computer response to button press. A standard card was always presented in all 3 positions along with the instruction to press (subjects always pressed the right-hand key for these trials). (d) Four cards were used for PCL trials in first and second sessions. Assignment of cards to subjects was pseudo-random.

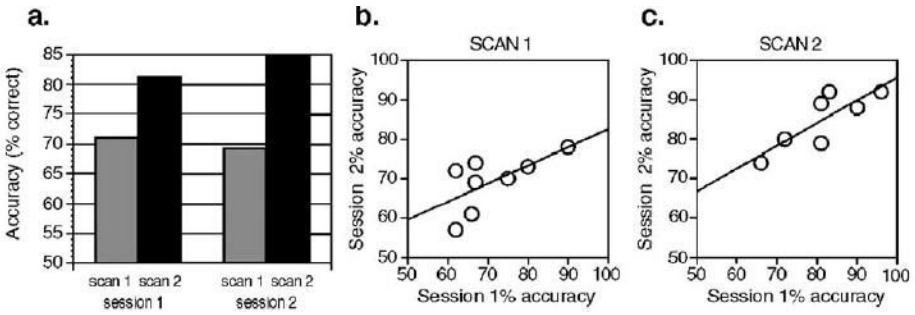

Fig. 2.

Behavioral data from first and second scanning sessions. (a) Mean accuracy for the subjects improved significantly across scans within each session ($P < 0.0001$), but there was no significant difference in accuracy between sessions. For the 8 subjects, mean accuracy for session 1 was significantly correlated with mean accuracy at session 2 (ICC = 0.85, $P < 0.01$), and this pattern was also evident for a between-session comparison of scan 1 (b) and scan 2 (c).

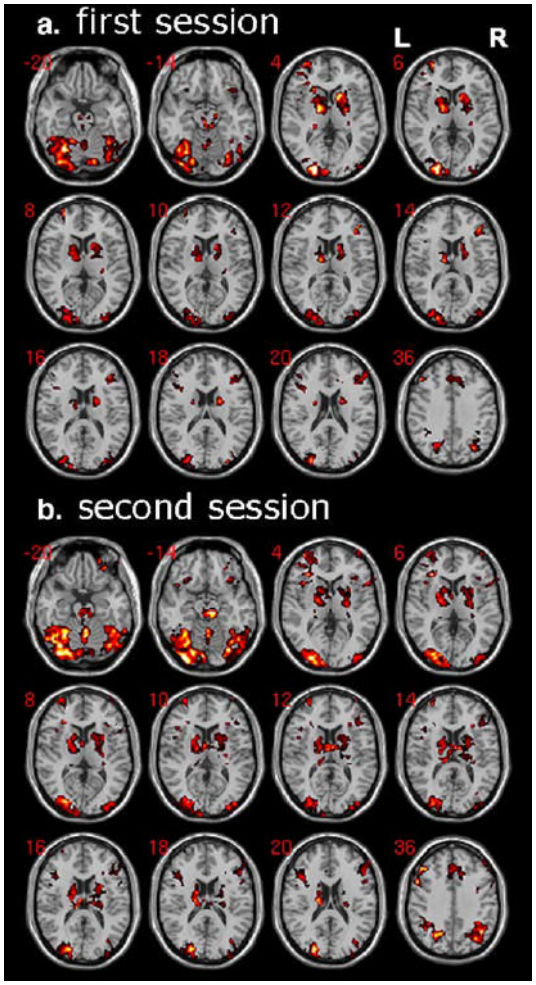

Fig. 3.

Learning in both sessions is associated with robust activation of midbrain and frontostriatal regions. For each session, a random effects analysis is run with contrast images (PCL trials minus baseline trials) for 8 subjects. The activations shown are significant after correction for multiple comparisons at the cluster level $P < 0.01$, voxel level threshold is $z > 2.3$. In both sessions, there is significant activation of midbrain, striatal, orbital, lateral and medial frontal cortex, as well extra-striate visual areas.

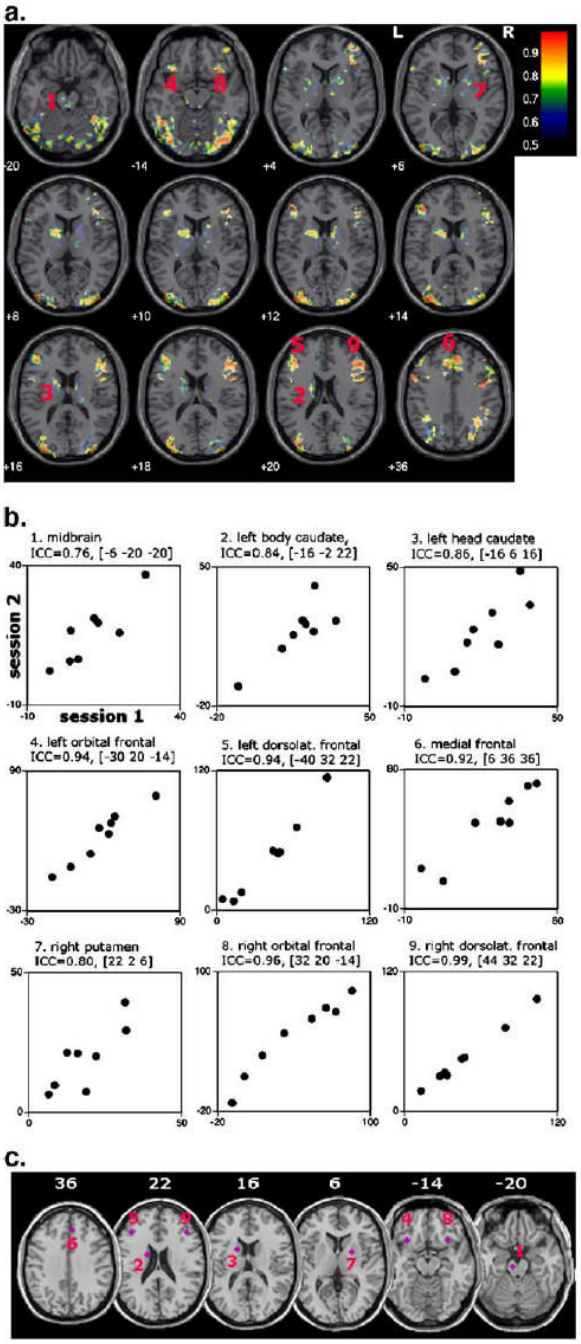

Fig. 4.

High test–retest reliability of fMRI signals within frontostriatal areas. (a) ICC values exceeding 0.5 are shown on a voxel-by-voxel basis within regions which were significantly activated for PCL vs. baseline for session 1 OR session 2 (inclusively). Voxels within midbrain, striatal, orbital, dorsolateral and medial frontal cortex show high ICC. (b) Illustrative signal plots within key regions of interest (ROIs) of this network. The ROIs were based on prior neuropsychological and neuroimaging research which has implicated midbrain, striatal and frontal foci (Aron et al., 2004; Knowlton et al., 1996; Moody et al., 2004; Poldrack et al., 2001; Seger and Cincotta, 2005; Shohamy et al., 2004). Mean signal within a sphere of 4 mm

radius was extracted for each subject and each session. The center of the sphere is demarcated by MNI coordinates [*x y z*]. (c) Panel showing each of the 9 ROIs on axial slices.

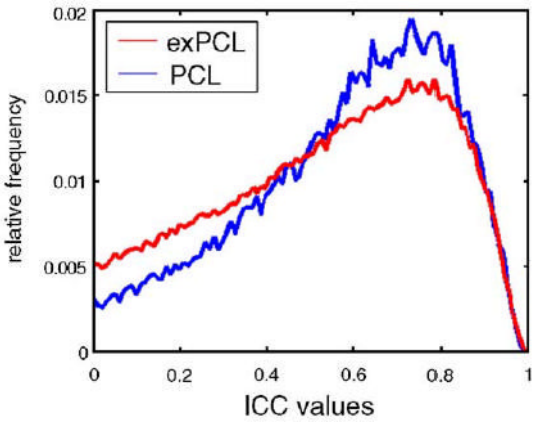

Fig. 5.

Reliability within the probabilistic classification learning (PCL) network is significantly higher than for brain regions outside that network. Relative frequency histogram of ICC values for PCL network and area outside PCL network, exPCL (excluding zero values and negative values in both cases). Values within the PCL network are significantly higher than for those in exPCL (Chi-square test for difference between distributions, $P < 0.0001$); confirming that test– retest reliability is greater in areas important for task performance.

Table 1

Complete information about stimuli for 100 trials

Card1	Card2	Card3	Card4	Stimulus	Frequency	Rain	p(rain)
1	0	0	0	1	7	1	0.14
0	1	0	0	2	7	1	0.14
0	0	1	0	3	7	5	0.71
0	0	0	1	4	7	4	0.57
1	1	0	0	5	8	0	0.00
1	0	1	0	6	12	11	0.92
1	0	0	1	7	1	1	1.00
0	1	1	0	8	7	1	0.14
0	1	0	1	9	1	1	1.00
0	0	1	1	10	19	18	0.95
1	1	1	0	11	19	6	0.29
1	0	1	1	12	2	1	0.50
1	1	0	1	13	3	2	0.67

Each of 13 stimuli consists of presentation of 1, 2 or 3 cards (the presence/absence of a card is indicated by 1/0 respectively). Each individual card is associated with the rain outcome across all 100 trials with the following probabilities: 0.18, 0.37, 0.59 and 0.82. The frequency of presentation of different stimuli ranges between 1 and 19. Each stimulus (consisting of 1, 2 or 3 cards), is associated, across the 100 trials, with varying probabilities of rain, ranging from 0 to 1.