Psychophysiology, 46 (2009), 367–378. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2008.00780.x

Beyond p-values: Averaged and reproducible evidence in fMRI experiments

MICHELLE LIOU,[a] HONG-REN SU,[a,b] ALEXANDER N. SAVOSTYANOV,[a,c] JUIN-DER LEE,[a] JOHN A. D. ASTON,[a,d] CHENG-HUNG CHUANG,[e] AND PHILIP E. CHENG[a]

[a]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[b]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[c]State Research Institute of Physiology, Siberian Branch of Russian Academy of Medical Sciences, Novosibirsk, Russia
[d]Department of Statistics, The University of Warwick, Coventry, UK
[e]Department of Computer Science and Information Engineering, Asia University, Taichung County, Taiwan

Abstract

In functional magnetic resonance imaging studies, there might exist activation regions routinely involved in experimental sessions, but modest in response magnitude. These regions may not be easily detectable by the conventional *p*-value approach using a rigid threshold. With particular reference to the reproducibility analysis method proposed in Liou and colleagues, this study presents some within- and between-subject brain-activation patterns that are replicable between experimental modalities, and robust to the method used for generating the patterns. There is a neurophysiological basis behind these reproducible patterns, and the conventional *p*-value approach using averaged data across subjects might not suggest the complete patterns. For example, recent studies based on the group-averaged data showed a task-induced deactivation in the precuneus and posterior cingulate, but our reproducibility analysis suggests both increased and decreased responses in the two regions. The increased responses localize in these regions with differentially distributed patterns for individual subjects and for different experimental tasks. In this study, we discuss the neurophysiological basis of the reproducible patterns and propose some applications of our research findings to scientific and clinical studies.

Descriptors: SPMs, Reproducibility, Category-preferential regions, Default network

Functional magnetic resonance imaging (fMRI) experiments are evidence-based inquiries; therefore, it is desirable to obtain as much information from the experimental data as possible. Research findings in fMRI studies are normally summarized using statistical parametric maps (SPMs), which highlight in an anatomical background those voxels exceeding a *p*-value threshold (e.g., $p < .05$). Brain voxels with smaller *p*-values are not just more responsive to experimental stimuli as compared with a control condition, they are also much greater in response magnitude. There might also be functional regions that are routinely involved in experimental sessions, but are modest in response magnitude—these regions could be easily bypassed in SPMs with a rigid threshold. Empirical studies have found that the functional regions locked to experimental tasks are typically responsive with smaller amplitude, but their responses are consistent throughout the experimental runs and subjects (e.g., Liou et al., 2003). On the other hand, image data collected in fMRI experiments are averaged across multiple runs and subjects in order to cumulate enough statistical power to detect activation regions. The final SPMs are constructed by applying the general linear model or other methods to the averaged data. Statistical analyses using group-averaged data have made a strong assumption that functional images after global normalization are interchangeable. There is a body of research showing that the degree of activation to experimental stimuli varies according to the stimulus sequence, effective strategy, and focused attention. The average of group data can possibly conceal more than it reveals, and negatively affects inference drawn from brain activation maps (Constable, Skudlarski, & Gore, 1995; Skudlarski, Constable, & Gore, 1999; Genovese, Lazar, & Nichols, 2002; Swallow, Braver, Snyder, Speer, & Zacks, 2003).

The SPMs constructed by the general linear model or other competing methods provide a useful summary of responses, and contribute one among many sources of evidence. Other supporting information beyond the average (or weighted average) could also lead to important insights into underlying cognitive functions. Reproducibility is defined as the number of experimental repetitions in which a voxel is consistently classified as active. In the literature, there have been methods proposed for investigating

The authors are indebted to Dr. Andreas Keil and an anonymous reviewer for helpful comments, and to the fMRIDC at Dartmouth College for supporting the data sets analyzed in this study. This research was supported by grant NSC96-2413-H-001-001-MY3 from the National Science Council (Taiwan).

Address reprint requests to: Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan. E-mail: mliou@stat.sinica.edu.tw

reproducible evidence without conducting separate fMRI experiments (e.g., Genovese, Noll, & Eddy, 1997; Liou et al., 2006; Strother et al., 2004). Specifically, the method proposed in Liou et al. (2006) was designed with a focus on both magnitude and persistency in brain activation using a random effect model. In the method, the threshold for assigning voxels to active/inactive status was decided for each individual subject according to image data in his/her in-brain voxels. The threshold was selected by maximizing the probability of making a correct decision among all other choices on the receiver-operator characteristic (ROC) curve. Unlike *p*-value thresholds, a threshold selected on the ROC curve would be less stringent for subjects whose images were more contaminated by noise. Because the method does not rely on the average of group data, it will provide useful information to supplement what is observed through SPMs. With particular reference to the method, we will present in this study some within- and between-subject patterns of reproducible evidence that are consistent between two experimental modalities, both involving visual stimuli in either object recognition or change detection tasks. We also demonstrate that functional differentiation between brain regions can be readily inferred from these reproducible patterns.

In the empirical study, we considered a data set that investigated the representation of objects in the human occipital and temporal regions through an on-and-off paradigm (Ishai, Ungerleider, Martin, Shouten, & Haxby, 1999; Ishai, Ungerleider, Martin, & Haxby, 2000). The data set was published by the US fMRI Data Center (fMRIDC), and has been reanalyzed in several studies (e.g., Liou et al., 2003; Mechelli, Gorno-Tempini, & Price, 2003). Because the design of experiments involved in the data set has optimized the interaction effects between stimuli, tasks, experimental runs, and subjects, it has been recommended as a benchmark for comparing between data processing methods (Liou et al., 2006). In the original SPM results, the ventral occipital and ventral temporal regions consistently showed differential responses to various objects. For example, the lateral fusiform gyri and inferior occipital gyri had greater responses to faces compared with other competing objects. The Ishai et al. (2000) study additionally found that category-related patterns of response were independent of tasks (passive viewing vs. delayed matching) and of spatial frequency differences in the stimuli (photographs vs. line drawings). In order to validate the patterns of reproducible findings between experimental modalities, we also included in the study another fMRIDC data set that was collected through an event-related paradigm for investigating brain functions in a change-detection task (Scott et al., 2001). In the original SPM results, the calcarine cortex was highly associated with task onset, and the dorsal and ventral visual regions were temporally associated with visual search. The study also found a network (e.g., inferior frontal cortical areas) associated with the execution of responses. Both studies involved visual stimuli, but there was no one-to-one correspondence between the design contrasts in the two data sets. For ease of exposition, we will mainly present reproducible patterns from the first data set, and use results from the second data set as a supplement to support reproducible findings between experimental modalities.

According to the dynamical localization model, physiological brain functions, such as regulation of breath or digestion, are identical across all subjects and have connections with specific brain areas (Vygotsky, 1931; Luria, 1964). But, higher mental functions are individual and their localization in the brain can be made only with some probability. Change of functional localization is an additional characteristic of a subject's psychological traits. The primary aim of this study is to demonstrate that there are within- and between-subject patterns of reproducible evidence, replicable between experimental modalities and robust to the particular method used for generating the evidence. There is a neurophysiological basis behind these reproducible patterns, and the conventional *p*-value approach using averaged data across subjects could have suggested incomplete patterns. For example, several SPM studies based on group-averaged data have shown a task-induced deactivation in the precuneus and posterior cingulate (e.g., Harrison, Yücel, Pujol, & Pantelis, 2007; Li, Yan, Bergquist, & Sinha, 2007), but our reproducibility analysis suggests both increased and decreased responses in the two regions. The increased responses in these regions have differentially distributed patterns for individual subjects and for different experimental tasks. Later in this study, we will discuss the neurophysiological basis of these reproducible patterns in more details. In the next section, the ROC thresholding method will be introduced in an intuitive approach. Based on the aforementioned data sets, we will present a summary of reproducible patterns between experimental modalities. According to the neurophysiological basis of these reproducible patterns, we finally propose some applications of our research findings to scientific and clinical studies.

Methods

Random Effect Models

In the SPM generalized linear model, the fMRI responses in the *i*-th run can be expressed as

$$(1)\ y_i = X_i\beta_i + e_i,$$

where y_i is the vector of image intensity after pre-whitening, X_i is the transformed design matrix, and β_i is the vector containing the unknown regression parameters. In (1), the pre-whitened data are assumed to have a Gaussian distribution with mean $X_i\beta_i$, and variance $\sigma_i^2 I_{n_i}$, where σ_i^2 is the residual variance associated with e_i, and I_{n_i} is the $(n_i \times n_i)$ identity matrix with n_i scan volumes in the *i*-th run. At the pre-whitening stage, spatially varying autocorrelations can be estimated by the residuals e_i after inserting design contrasts and all known MRI artifacts (as covariates) into the design matrix (cf. Worsley et al., 2002, equation 4). Without knowing imaging artifacts well, however, the size of autocorrelations can be over-determined and -corrected. The SPM software supplies a few common artifacts for preprocessing functional images. These common artifacts can be considered in data analyses.

The use of random effect models has been proposed for finding a weighted average of image data across subjects (cf. Friston et al., 2002; Worsley et al., 2002). If image data are not interchangeable between runs and between subjects, estimates of experimental effects are preferably obtained from individual runs corresponding to each subject separately (Constable et al., 1995; Skudlarski et al., 1999). In the random effect model, the regression parameters β_i are additionally assumed to be random from a multivariate Gaussian distribution with common mean μ and variance Ω. The empirical Bayes estimate of β_i in the random effect model shrinks all estimates toward the mean μ, with greater shrinkage at noisy runs (i.e., greater σ_i^2). Because of the shrinkage, the values of estimates become closer to each other and,

therefore, the reproducibility between runs can be optimized. The size of μ can further be modeled as between-run design effects locked to experimental tasks. We suggest considering stimulus effects within the design matrix X_i in (1), and task effects within the design matrix associated with μ. Analogous to the general linear model, t-values of a particular contrast within runs can be computed by normalizing the estimated $\hat{\beta}_i$ using the corresponding standard error $\hat{\sigma}_i$. For each design contrast, there are M such t-values and M is the total number of runs. The T-values due to task effects can be found by the same analogy using parameter estimates of μ and Ω. If there is no interaction effect between stimuli and tasks, the T-value of a voxel is simply an average of overall effect across runs attributable to a contrast, for example, faces versus houses/chairs.

Thresholding

In fMRI studies, the true status of each voxel is unknown but can be estimated using the t-values within runs derived from the random effect model. If we select K distinct thresholds in increasing order of magnitude, these M t-values (in absolute value) can be classified into $K+1$ groups. Let P_{A_k} denote the conditional probability of t-values assigned to the k-th group given the truly active status, and P_{I_k} carries the same definition, but given the truly inactive status. Both P_{A_k} and P_{I_k} are unknown with observed counts γ_k. If a particular threshold k^* is selected, sensitivity and the false alarm rate are defined respectively as

$$(2)\quad \begin{aligned} \text{Sensitivity} &\equiv P_A = \sum_{k=k^*}^{K} P_{A_k}, \quad \text{and} \\ \text{False Alarm} &\equiv P_I = \sum_{k=k^*}^{K} P_{I_k}. \end{aligned}$$

The ROC curve is a bivariate plot of P_A versus P_I given all possible thresholds. For a total of V in-brain voxels, these unknown P_{A_k} and P_{I_k} can be estimated by assuming a mixed multinomial distribution in the following likelihood:

$$(3) L(P_A, P_I, \lambda | r) = \prod_{v}^{V} \left[\lambda \prod_{k}^{K} P_{A_k}^{r_k^{(v)}} + (1-\lambda) \prod_{k}^{K} P_{I_k}^{r_k^{(v)}} \right] I(\lambda \le \rho),$$

where r denotes a collection of observed counts, and $r_k^{(v)}$is the observed count in the k-th group of the v-th voxel. In the likelihood, a prior $I(\lambda \le \rho)$ is employed, where I(E) is the indicator function of the event E, and ρ is a prior of the unknown λ parameter for the proportion of active voxels.

By maximizing $L(P_A,P_I,\lambda|r)$, we can obtain K pairs of (P_A, P_I), and the ROC can be interpolated via a smoothed function. In fMRI applications, we need to assign voxels an active/inactive status with a decision threshold. The reproducibility of a voxel is defined as the degree to which the "active status" of a voxel, in responding to stimuli, remains the same across experimental runs. Given a decision threshold k^* on the observed t-values, the estimated true status and classification results can be organized into a 2×2 table. Let the proportion of correct classification be P_o in the table, and its expected value is P_c. In the literature, the Kappa index (Cohen, 1960) and Type-I error (or false alarm) are defined respectively as

$$(4)\quad \begin{aligned} \text{Kappa} &= \frac{P_o - P_c}{1 - P_c}, \\ \text{Type} - \text{I error} &= P_I. \end{aligned}$$

In this study, we suggest selecting the decision threshold k* which maximizes the Kappa value.

SPMs and Reproducible Evidence

In the empirical study, the decision threshold was selected by maximizing the Kappa value. By convention, we categorized voxels according to reproducibility (i.e., a voxel is strongly reproducible if its active status remains the same in at least 90% of the runs, moderately reproducible in 70–90% of the runs, weakly reproducible in 50–70% of the runs, and otherwise not reproducible), and the brain activation maps were constructed on the basis of strongly reproducible voxels. In order to take into account image distortion due to slice timing and motion correction, the brain maps also included voxels that were moderately reproducible and spatially proximal (nearest neighbors) to strongly reproducible voxels.

Empirical Examples

Experimental Data

In this study we mainly considered the data set that involved twelve subjects, each performing twelve runs of either delayed match-to-sample or passive viewing tasks (Ishai et al., 2000). In the delayed matching task, a target stimulus was followed, after a 0.5-s delay, by a pair of choice stimuli presented at a rate of 2 s. Subjects indicated which choice stimuli matched the target by pressing a button with the right or left thumb. In the passive viewing task, a stimulus (house, face, or chair) was presented at a rate of 2 s and subjects simply responded to stimuli without recording a target or making a decision on choice stimuli. All runs involved phased, scrambled pictures presented at the same rate as the control stimuli. There were three orthogonal contrasts examined in the experiments—namely, meaningful objects (i.e., faces, houses, and chairs) versus the control condition (i.e., phased, scrambled pictures), faces versus houses/chairs, and houses versus chairs (Ishai et al., 1999; Ishai et al., 2000). In this study, we inserted the same orthogonal contrasts into the design matrix in (1) without convolution of any hemodynamic response function.

In order to validate the results from the first data set, we also included an additional ten subjects who participated in a change-detection task of 10–12 runs, each consisting of 10 stimulus trials (Scott et al., 2001). In each trial, there were two images in a pair with difference in either the presence/absence of a single object or the color of the object. The subjects made behavioral responses by pressing a button when they felt that there was something changing on the trial. For the first 30 s of the trial, the two images were presented for 300 ms, separated by a 100-ms mask. The mask was removed during the last 10 s of the trial, and the stimuli alternated every 400 ms. In the original analysis, the hypothetical waveforms were specified according to task onset, visual search, response execution, and deactivation. Because those hypothetical waveforms might not be orthogonal to each other, we inserted Laguerre polynomials of the first and third orders into the design matrix in (1) (Saha et al., 2004; Su, Liou, Cheng, Aston, & Lai, 2007). The two contrasts examined brain responses that were continued within the 40-s trial, and responses that were different between the stimulus presentation with and without the mask, respectively.

Decision Statistics

In the data analysis, the effects due to different contrasts for each run along with the average effect across runs were computed using the random effect model for each individual subject. Given an optimal threshold, the number of reproducible runs were

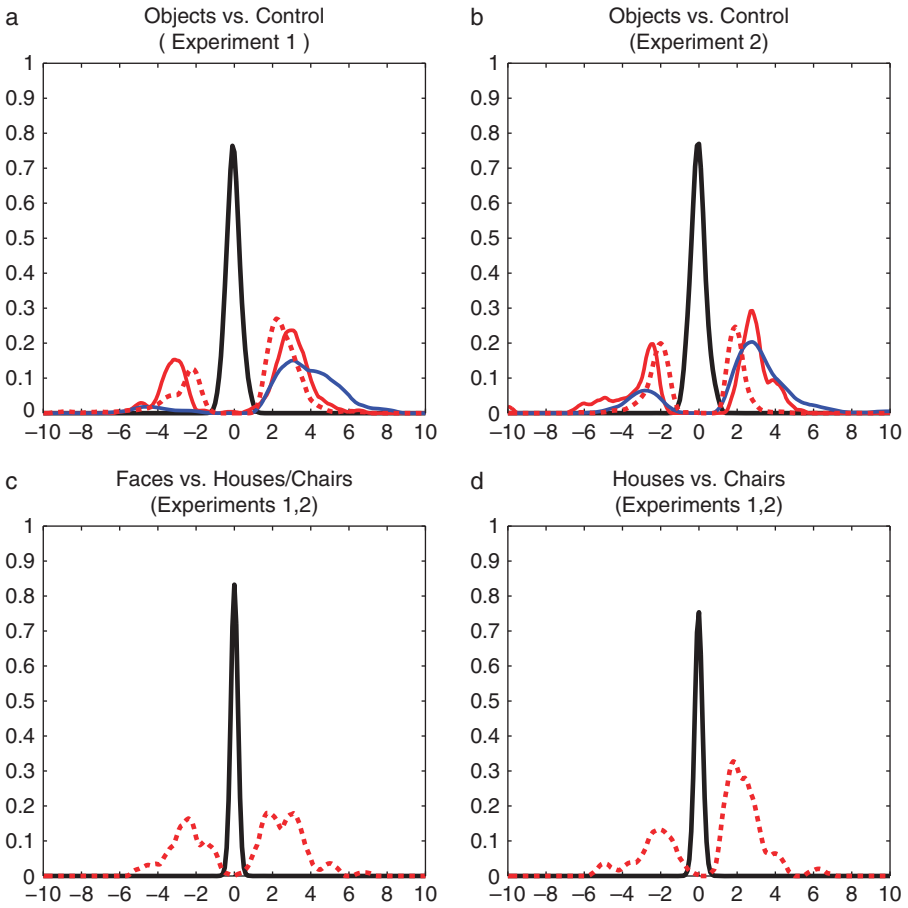

Figure 1. The density distributions of average *T*-values across subjects for different contrasts in Experiments 1 and 2. The areas under different distributions are normalized to have the same value of one. The average values for voxels consistently classified as inactive in the 12 runs are plotted in black; those consistently classified as active in the 12 runs are plotted in red; voxels classified as moderately reproducible in the 12 runs (i.e., 8 to 10 runs) are plotted as dotted line in the figures. Those strongly reproducible voxels located in the precuneus with either positive or negative *T*-values are plotted in blue. Subjects in Experiment 1 participated in passive viewing and delayed matching tasks. Because negative *T*-values in the precuneus are mainly locked to the delayed matching task, the blue plot in the lower tail in (a) is less compelling as compared with the same plot for Experiment 2 in (b), which involved both delayed matching of photographs and line drawings. Across all subjects, there is no voxel consistently more active to faces compared with houses or chairs in the 12 runs, nor is any voxel more active to houses compared with chairs. Therefore, the plots in (c) and (d) only give distributions of *T*-values that are moderately reproducible.

computed using the *t*-values within runs. According to the empirical results, the Kappa values are higher when comparing meaningful objects with the control condition for the Ishai et al. (1999) study. This means that image data collected in the experiments are more reliable for examining this contrast. Also, empirical Type-I errors (i.e., the false alarm values in (2)) between subjects range from 0.03 to 0.06 for this contrast. When comparing between objects (faces versus houses/chairs or houses versus chairs), however, the Kappa values are reduced to a range of 0.25 to 0.33, and Type-I errors slightly increase to 0.08–0.10. For illustration, distributions of the average *T*-values across subjects are plotted in Figure 1 for different contrasts in the Ishai et al. (1999) study. The average *T*-values are plotted separately for strongly and moderately reproducible voxels. As a comparison, we also plot the average values for those voxels consistently classified as inactive across runs. It is interesting to note that strongly (active in at least 11–12 runs) and moderately (active in at least 8–10 runs) reproducible voxels have a sizeable overlap in their *T*-values. The plots suggest that, on average, the magnitude of *T*-values does not directly imply reproducibility. The decision statistics for the Scott et al. (2001) study suggest a similar pattern. The ROC curve was estimated using the likelihood in (3), which might give biased estimates of λ and other conditional probabilities. The empirical Kappa values and decision errors could have been changed when the true parameter values were known. However, the relative sizes of Type-I errors across design contrasts should remain the same regardless of bias in estimates. It is therefore interesting to note that the Type-I errors for each design contrast differ only within a range of .01 to .02, even though images of individual subjects are noisy to a greater or lesser degree.

Results of Comparing Objects with the Control Condition

The *T*-values for those strongly reproducible voxels in the precuneus are also plotted in Figure 1 for the Ishai et al. (2000) study. The decreased activities in the precuneus are mainly associated with the delayed match-to-sample task. Experiment 1 involved only six runs of the delayed matching task. Therefore, the distribution in the negative tail is almost invisible for the six subjects in Experiment 1. It is clear that the *T*-values of strongly

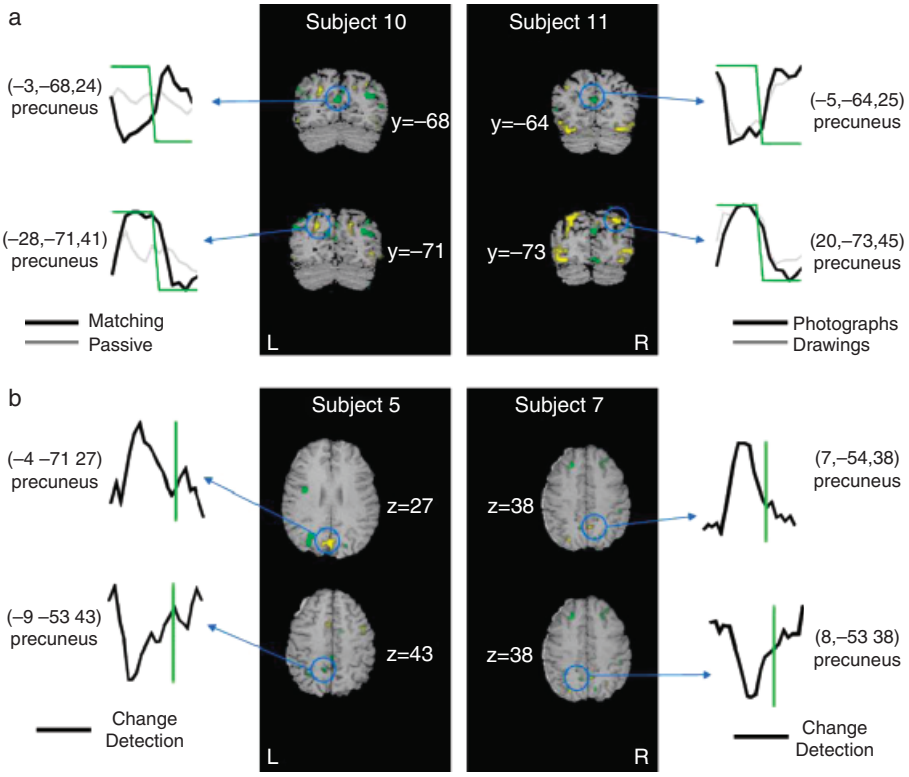

Figure 2. The HRFs and activation maps in the precuneus for (a) 2 subjects in the Ishai et al. study (2000) and (b) another 2 subjects in the Scott et al. (2001) study. Subject 5 in the Scott et al. study had longer reaction time as compared with Subject 7. Coordinates are in the normalized space of the Talairach and Tournoux 1988 brain atlas. The HRFs corresponding to different regions are the averages of observed images across stimuli and runs without any normalization except for a mean shift such that different functions can be shown in the same graph. Because the decreased activities are only specific to the delayed matching and change detection tasks, the HRFs are plotted separately for the passive viewing and delayed matching tasks for Subject 10 in the Ishai et al. study.

reproducible voxels in the precuneus are also widely distributed. Subjects engaged in the change detection task also consistently show positive and negative responses in the precuneus. Figure 2 gives the activation maps in the precuneus for different subjects in the two data sets. The maps were constructed using mri3Dx (http://www.idoimaging.com), which automatically performed three-dimensional image rendering. For the two subjects engaged in the matching tasks, the medial regions of the precuneus have decreased activity, but the decreases are spatially closer to the increases in the lateral site of the precuneus. The increased/decreased pairs are spatially distributed in the medial and lateral regions of the precuneus for all subjects in the Ishai et al. study (Subject 2 has no clearly decreased responses in the precuneus). The two subjects engaged in the change detection task, however, do not follow the same activation patterns. The increased/decreased pairs for the two subjects are distributed closer to the medial site of the precuneus. The hemodynamic response functions (HRFs) in Figure 2 were computed using the averaged responses across trials and runs. The HRFs suggest that the increased/decreased responses are restricted to the delayed matching and change detection tasks. The passive viewing task involves only increased responses in the precuneus. More specifically, for the six subjects who participated in six runs of the passive viewing and six runs of the delayed matching tasks, there are three types of activation patterns in the precuneus, namely, regions with increased responses in both passive viewing and delayed matching, regions with no clear responses in passive viewing but with decreased responses in delayed matching, and regions with no clear responses in passive viewing but with increased responses in delayed matching.

Neuroimaging research has recently suggested that specific human brain areas are tonically active in a resting state and deactivated when subjects are engaged in a wide variety of cognitive tasks (Raichle et al., 2001; Shulman et al., 2002). The precuneus showing increased and decreased activities are located in the default areas and specific to the delayed matching and change detection tasks. The physiological mechanisms behind the decreases in the default areas are still under investigation. Thus far, supporting evidence has posited that some of the decreases that are observed in areas remote from activations possibly reflect the inhibition of information processing in areas that are not engaged in task performance (Gusnard & Raichle, 2001). In Table 1, we list a few regions that simultaneously show increased and decreased responses in the two data sets. According to the table, the two studies involve the increased/decreased responses in the lingual gyrus, cuneus, precuneus, and posterior cingulate. The lingual gyrus shows positive and negative responses in all types of tasks (i.e., passive viewing, delayed matching, and change detection). We will discuss the neurophysiological basis of those reproducible patterns later.

Results of Comparing Between Objects

In the two experiments of the Ishai et al. (2000) study, there is no strongly reproducible voxel that consistently shows increased responses to faces relative to other objects throughout the twelve runs, nor does any voxel consistently show greater activity to

Table 1. *Reproducible Regions Showing Increased/Decreased Responses*

The Matching Task (Ishai et al., 2000)												
	Experiment 1						Experiment 2					
Subjects	4	6	7	9	10	12	1	2	3	5	8	11
Ling. gyrus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−
Cuneus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−
Precuneus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	−	+/−	+/−	+/−	+/−
Post. cing.	+/−	+	+/−	+/−	+/−	+	+/−	+/−		+/−		−

The Change-Detection Task (Scott et al., 2001)										
Subjects	1	2	3	4	5	6	7	8	9	10
Ling. gyrus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−
Cuneus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−
Precuneus	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−	+/−
Post. cing.	+/−		+/−	+/−	+/−		+/−		+/−	+/−

either houses or chairs. The results are not surprising because recent research on decoding of mental states suggested 70% to 80% accuracy that could be achieved when predicting category-related responses without knowledge of what a subject perceived in the fMRI experiment (Haynes & Rees, 2006). For illustration, Figure 1 also gives the distribution of *T*-values for moderately reproducible voxels when comparing between objects. Because the category-related patterns of response are independent of tasks in the Ishai et al. study and also in our data analyses, the distributions are plotted for the average *T*-values across the twelve subjects. In order to investigate if responses in the non-reproducible runs could exhibit a random or systematic pattern, we selected a few moderately reproducible voxels in the posterior cingulate and inferior temporal regions, and looked into the non-reproducible runs for two subjects having higher Kappa values as compared with other subjects in the two experiments (i.e., Subjects 3 and 6).

Subject 3 participated in six matching of drawings (D) runs followed by six matching of photographs (P) runs (i.e., D3 → D5 → D1 → **D4** → D6 → **D2** → P3 → P5 → P1 → **P4** → P6 → **P2**). In the original design of experiments, the stimuli had a distinct order of presentation in each run for balancing the sequence effect. For example, the D2 run was designed to give faces ahead of houses (i.e., Faces → Chairs → Houses → Faces → Houses → Chairs), and D4 to give houses first (i.e., Houses → Chairs → Faces → Houses → Faces → Chairs). For this subject, those moderately reproducible voxels showing greater activity to faces in other runs were more active to houses in D2 and D4 (also in P2 and P4). If the design sequence was D3 → D5 → D1 → **D2** → D6 → **D4** → P3 → P5 → P1 → **P2** → P6 → **P4**, those moderately reproducible voxels should have been strongly reproducible for comparing between faces and houses/chairs in the random effect model (note: D2/P2 and D4/P4 are swapped in their relative positions in the design sequence). Similarly, the voxels of greater activity to houses were more active to chairs in D4/P4 and D6/P6, which were designed to give either houses or chairs ahead of other stimuli (cf. the fMRIDC document Accession No. 2-2000-1113D). Subject 6 participated in six matching of photographs (P) runs and six passive viewing (V) runs. For this subject, brain regions showing greater activity to faces were reproducible in all runs except for V1, V5, P2, and P4 in which the same regions had increased responses to either houses or chairs. For the same subject, regions of greater activity to houses were more active to chairs in V3, V5, P4, and P6.

Before reanalyzing the fMRI data, we created ad hoc sequences for Subjects 3 and 6 in which a few non-reproducible runs were swapped in their relative positions in the original experimental protocol. Image data were analyzed again by inserting the ad hoc sequences into the random effect model and pretending that the category-related responses followed the new sequences of stimulus presentation. Figures 3 and 4 give activation maps of category-preferential regions using the ad hoc sequences for Subjects 3 and 6, respectively. For ease of comparison, the activation maps are shown at the same positions on image slices (comparable y-coordinates) for both subjects. The HRFs in the figures were computed by matching the "original image data" to the ad hoc sequence rather than to the original design protocol.

As shown in the figures, the face-preferential regions are equally distributed in the parahippocampal gyrus, posterior cingulate and fusiform gyrus because the aforementioned regions were all moderately reproducible in the original experiments. According to the neural model of stimuli, an estimation of information novelty is connected with neuron cells localized in the hippocampus and parahippocampal gyrus (Mangina & Sokolov, 2006; Sokolov, Nezlina, Polyanskii, & Evtikhin, 2002). Such reaction is unspecific to any stimulus modality or class of tasks. The response magnitude in the parahippocampal gyrus depends not only on the category of stimuli, but also on the order of stimulus presentation. In other words, the parahippocampal gyrus and fusiform gyrus could be functionally different, even though their activation patterns to different categories of stimuli are almost identical. As a comparison, Figure 5 gives the activation patterns in the parahippocampal gyrus and fusiform gyrus for Subject 7 in the Scott et al. (2001) study. In the change detection task, the parahippocampal gyrus is clearly involved with both positive and negative responses, but the fusiform gyrus is only involved with positive responses that last longer within the 40-s trial. Our findings unnecessarily cast doubt on category-preferential responses. Instead, the reproducible evidence suggests that response magnitude may not be directly interpretable without comparing observations across experimental modalities.

In the original data analysis, only limited voxels in the posterior cingulate and inferior temporal regions showed moderately

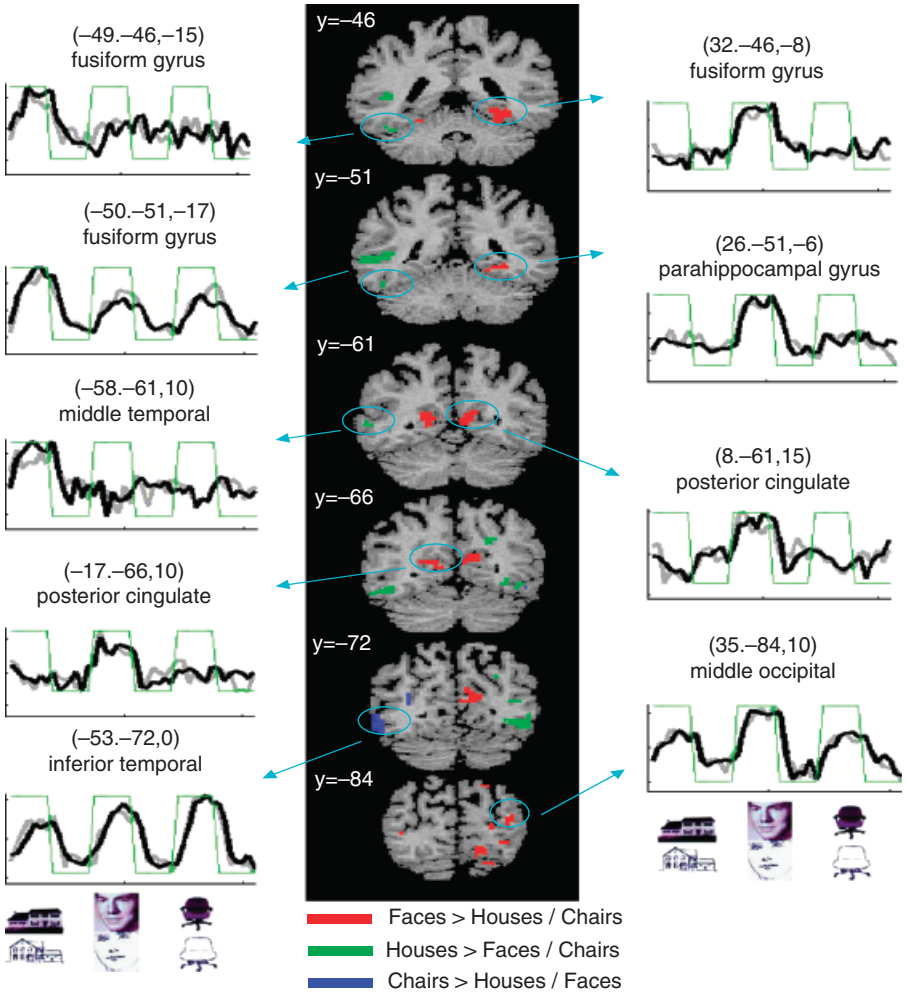

Figure 3. The brain activation maps for category-preferential regions constructed by replacing the original design sequence with the ad hoc sequence chosen for Subject 3. The maps are constructed based on the ad hoc sequence: D3 → D5 → D1 → D2 → D4 → D6 → P3 → P5 → P1 → P2 → P4 → P6. The colored voxels were moderately reproducible in the original analysis. The HRFs corresponding to different regions are the averages of observed images across stimuli and runs according to the ad hoc sequence, without any normalization except for a mean shift such that different functions can be shown in the same graph. The darker line in each graph is for delayed matching of photographs and the lighter line is for delayed matching of line drawings.

reproducible patterns. Because of the ad hoc sequences, experimental effects became noticeable via the random effect model for those colored regions in Figures 3 and 4. In order to verify that all those colored regions in Figures 3 and 4 were moderately reproducible, Figures 6 and 7 give the HRFs of the averaged responses over all runs and those over non-reproducible runs separately for category-preferential regions computed according to the original design protocol. We notice that a brain region preferential to a category of stimuli on average could have shown greater activity to competing objects one time out of three in the same experiment. The results also suggest that the particular methodology used for generating reproducible evidence should have found those strongly reproducible regions if the category-related effects were indeed reproducible across runs.

Discussion and Conclusion

The methodology used for finding reproducible patterns in this study has been designed to maximize the between-run reproducibility via the random effect model. Although the threshold selected by maximizing the Kappa value may control the empirical Type-I error within a reasonable range, there must be a sizable number of false positive hits among those voxels being classified as active within each run. By counting on the strongly reproducible criterion, the method may still preserve enough true positive voxels and bypass those false positives. A complete analysis of the experimental data in the Ishai et al. (2000) and Scott et al. (2001) studies suggests that the Kappa index along with the reproducibility criterion offer findings beyond those obtained by the SPM approach. The 22 subjects in the two data sets consistently show a pattern of increased/decreased responses in the precuneus, regardless of a wide range of Kappa values associated with individual subjects. But, the increased/decreased patterns in the precuneus are observable when subjects perform the delayed matching and change detection tasks. The increased/decreased pairs have been observed in other regions such as the lingual gyrus, cuneus, and posterior cingulate as well.

In the neurophysiological literature, there is a distinction between perception of objective or physical parameters and perception of subjective or psychological parameters of visual stimuli (e.g., Ivanitskii, Strelez, & Korsakov, 1984; Ivanitskii,

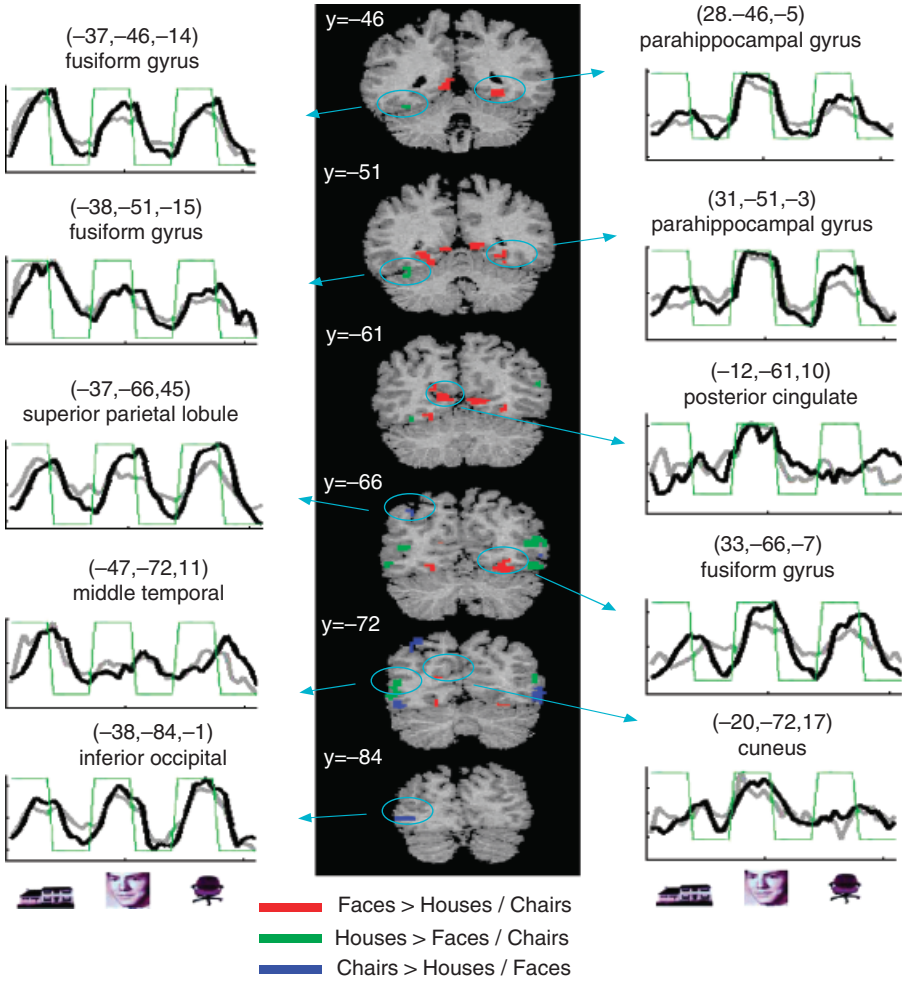

Figure 4. The brain activation maps for category-preferential regions constructed by the ad hoc sequence chosen for Subject 6. The maps are constructed based on the ad hoc sequence: P3 → V5 → P5 → V4 → P1 → V6 → P6 → V2 → P2 → V1 → P4 → V3. The colored voxels were moderately reproducible in the original analysis. The HRFs corresponding to different regions are the averages of observed images across stimuli and runs according to the ad hoc sequence, without any normalization except for a mean shift such that different functions can be shown in the same graph. The darker line in each graph is for delayed matching of photographs and the lighter line is for passive viewing of photographs.

1996). Objective parameters refer to the color, size, form, spatial location (or distance), and speed of movement of the stimuli. Those parameters are estimated by the "specific" sensorial areas in the cortex. The subjective parameters, on the other hand, refer to the emotionality, novelty, and importance of the perceived images in an attention-demanding task. Those parameters are estimated by the "unspecific" sensorial areas in the cortex with possible connection to the limbic system. According to Ivanitskii et al. (1984), the lingual gyrus is one of the specific visual areas where activity depends on the physical parameters of the images rather than on the psychological parameters of the subjects. Different activation patterns inside this area result from participation of neuron cells in perception of different physical parameters. The groups of neuron cells responsible for perception of relevant signals may inhibit an activity of the neighboring neuron cells responsible for the perception of irrelevant signals, the so-called lateral inhibition, to sharpen the spatial profile of excitation in response to a localized stimulus (e.g., Blakemore & Tobin, 1972). The differentiation in the lingual gyrus depends on the physical parameters of the visual stimulus and is independent of experimental modalities. This could explain why the positive/negative responses in the lingual gyrus are strongly reproducible between subjects and experimental tasks.

It is also true that the experimental tasks considered in this study involve different degrees of complexity. As compared with the passive viewing task, for example, delayed matching and change-detection require not only stimulus perception, but also evaluation and comparison between stimuli for decision making. Brain functions such as memory will be necessary in complicated tasks. Results of our data analysis show that more complicated tasks induce not only sharper response functions, but also more focused attention (or stronger selectivity in spatial and motivated attention; cf. Keil, Moratti, Sabatinelli, Bradley, & Lang, 2005; Lang, Bradley, & Cuthbert, 1997). Selective attention is involved in simple tasks such as passive viewing as well, because the perception of physical parameters requires subjects to distinguish an object from its background. As tasks become more complicated, however, differentiation in the lingual gyrus is not enough, and additional fields such as the cuneus and precuneus also participate in information processing. In the cuneus and precuneus, there are cell groups responsible for focused attention to different parts of the visual space. In the execution of matching and change

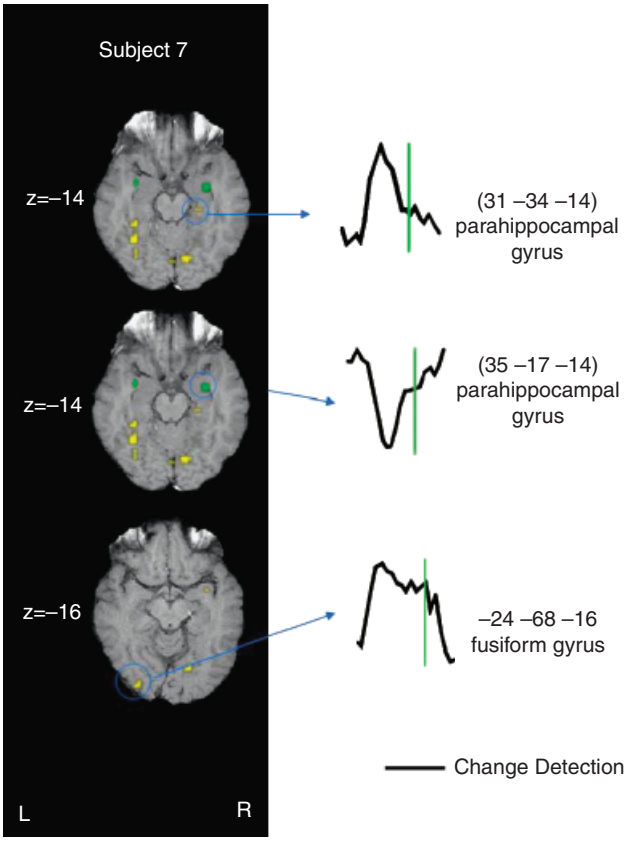

Figure 5. The activation regions in the parahippocampal gyrus and fusiform gyrus for Subject 7 in the Scott et al. (2001) study.

detection tasks, there are relevant and irrelevant fields in the visual space. Spatially selective attention in complicated tasks could be a reason for finding positive (relevant visual space) and negative (irrelevant visual space) responses in the two regions when subjects perform the matching and change detection tasks. Subject 4 in the Ishai et al. (2000) study engaged focused attention in the passive viewing task and showed decreased activity in the precuneus.

Spatially selective attention is also connected with fixation of eye positions at some points, and with inhibition of eye movement to irrelevant spatial fields. Complicated tasks need more intensive eye movements in image recognition. According to Olson, Musil, and Goldberg (1996), the posterior cingulate participates in eye movements in visual-motor tasks. The neurons in this region fired during periods of ocular fixation at a rate determined by the angle of gaze and by the size and direction of the preceding eye movement. Several parts of the posterior cingulate are connected with attention-related eye fixation at different positions. In our study, positive responses are likely observable when a subject's eyes are fixed at a particular part of the stimulus, and otherwise unobservable when the subject is constantly moving his/her eyes without fixation. This interpretation is tentative because there is no data showing eye positions in the two data sets.

A change in stimulus sequences may cause a distinction in novelty of the perceived images. The stimulus, which is perceived as new as compared with previous ones, results in the activation of the unspecific brain structure especially connected with the limbic system (Tulving, Markowitsch, Kapur, Habib, & Houle, 1994), which is a center of motivated attention. Our results suggest that the motivated attention might be weaker in delayed matching and change-detection than in processing emotional pictures (Keil et al., 2005; Lang et al., 1997). Motivated attention strongly varies between subjects because perceiving a stimulus as new or old depends on the psychological parameters of a subject. The between-subject reproducibility of the positive/negative responses is reduced in the parahippocampal gyrus in the two data sets (i.e., two subjects in the Ishai et al. study and four subjects in the Scott et al. study show positive and negative responses in the parahippocampal gyrus).

Recent SPM studies using the group-averaged responses have consistently reported a task-induced deactivation in the precuneus and posterior cingulate (e.g., Li et al., 2007; Harrison et al., 2007). However, our reproducibility analysis based on the data of each individual subject suggests both positive and negative responses in the two regions. During the experiments, there could be fields of visual space that were always irrelevant to experimental tasks. The selective attention of subjects should not be directed to this visual field, and could be consistently inhibited in all subjects to disallow irrelevant reactions. Because of the irrelevant space with unchanged boundaries, the decreased responses were consistent between subjects and easily detectable by analysis of group-averaged data. During object recognition, however, different parts of the stimuli in the visual field on which attention was directed varied between subjects, and increased responses in the unspecific sensorial areas might not be easily found in the averaged data. Because the irrelevant space also inhibited eye fixation at some positions, negative responses in the posterior cingulate are more likely to have a similar interpretation as those in the precuneus.

By comparing between objects of different categories, our results suggest that category-preferential regions are not strongly reproducible between experimental runs; this is true for all subjects in the Ishai et al. (1999) study. According to the HRFs in Figures 6 and 7, a region preferential to faces can be more responsive to houses one time out of three and still show a greater response to faces on average. The object is perceived visually whereas the category of objects is an abstract concept defined by verbal constructions. Our reproducibility analysis only examines if a particular region can be consistently more responsive to a category of stimuli, and the answer is "no." When performing experimental tasks, subjects name an object using internal speech and press the button according to the name. It is possible to assume that the brain area that is responsible for object recognition is defined by a subject's internal speech. As the choice of a category concerns the highest mental functions, its localization can only have a probable basis and vary from one stimulus to the other. In the literature, there has been a complete discussion on the roles of focused attention, anxiety, cognitive state, viewing angles, and other extraneous factors in generating category-preferential responses (Wojciulik, Kanwisher, & Driver, 1998; Gauthier, Skudlarski, Gore, & Anderson, 2000; de Gelder & Rouw, 2001; Joseph & Gathers, 2002). The HRFs in Figures 6 and 7 suggest that an extensive use of the group-averaged responses may overlook important evidence about the functional architecture in the ventral occipital and ventral temporal regions. The ad hoc sequences designed for Subjects 3 and 6 indicate that there is still a possibility to control confounding effects, if known, and obtain strongly reproducible comparisons between objects.

The signal to noise ratio is greater for the on-and-off paradigm than that for the event-related paradigm. In the two data sets, the fusiform gyrus is involved only with positive responses,

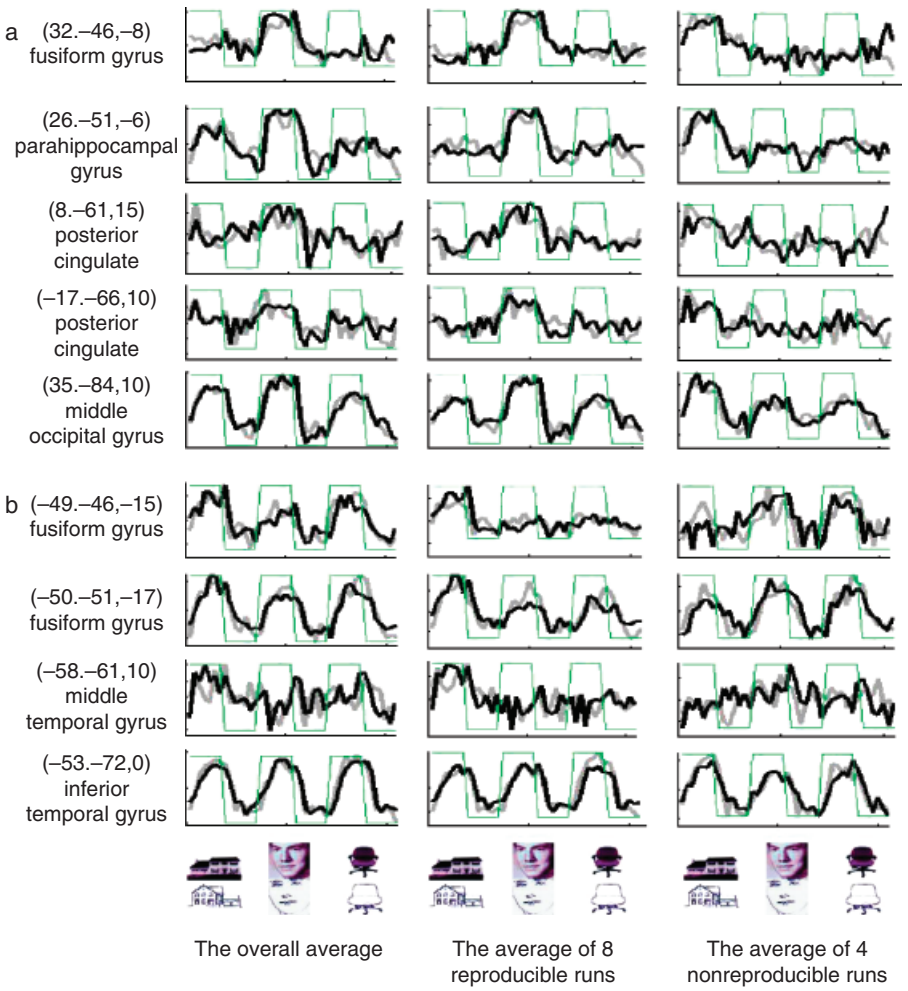

Figure 6. The HRFs of category-preferential regions for Subject 3. The HRFs corresponding to different regions are the averages of images across runs according to the original design sequence: D3 → D5 → D1 → D4 → D6 → D2 → P3 → P5 → P1 → P4 → P6 → P2. The four non-reproducible runs in (a) refer to D2, D4, P2, and P4, and in (b), refer to D4, D6, P4, and P6.

and this observation is strongly reproducible between subjects and between experimental modalities. Other regions such as the inferior occipital gyrus and inferior temporal gyrus are also involved with only positive responses and are strongly reproducible between subjects. There are positively responding regions that are not reproducible between subjects, but highly correlated with behavioral data; for example, the superior frontal gyrus and supramarginal gyrus are positively active and strongly reproducible only for subjects with longer reaction time in the Scott et al. (2001) study. In this study, we have carefully verified the methodology used for generating reproducible patterns of responses. Based on the reproducible patterns in the two data sets, we have also found that brain responses due to information synthesis (Ivanitsky, 1996) are likely localized in brain areas that are only involved with increased activities. However, perceptual responses more likely involve increased and decreased activities localized in brain regions into which there are functional differentiations. This observation is only tentative, and an interested reader may consider our research findings preliminary work toward the integration of reproducible patterns between experimental modalities.

Our data analysis suggests that a few brain regions consistently show positive and negative responses in different experiments involving visual stimuli. These research findings are in agreement with the Ivanitskii hypothesis (1996) of two neurophysiological aspects in stimulus perception—specific and unspecific sensorial structures. The lingual gyrus is a part of the specific structure, and its activation patterns are likely determined by the physical parameters of the stimulus or by selective attention in simple tasks. The cuneus, precuneus, and posterior cingulate are connected with the unspecific structure, and the increased/decreased activities in these regions likely reflect selective attention in complicated tasks or the psychological parameters of subjects. In clinical applications, it is meaningful to separate perceptual dysfunctions, such as sensory sensitivity, from those with underlying psychological grounds. The reproducible patterns in sensorial brain areas may provide a guideline to clinical diagnosis of perceptual dysfunctions. Research findings in this study also support the hypothesis that focused attention is achieved by activation of some, and inhibition of other, neuronal cells. The increased/decreased patterns are evidence of functional separation between different neuronal cells. When a task is simple (e.g., passive viewing), functional separation disappears in the unspecific structure. These results may be applied to studies on the process of intra-regional separation in the cell activity. For example, according to the Lubow (1989) hypothesis, attentional deficit in schizophrenia results from hyperactivity of neuronal cells and disruption of inhibition processes, whereas the

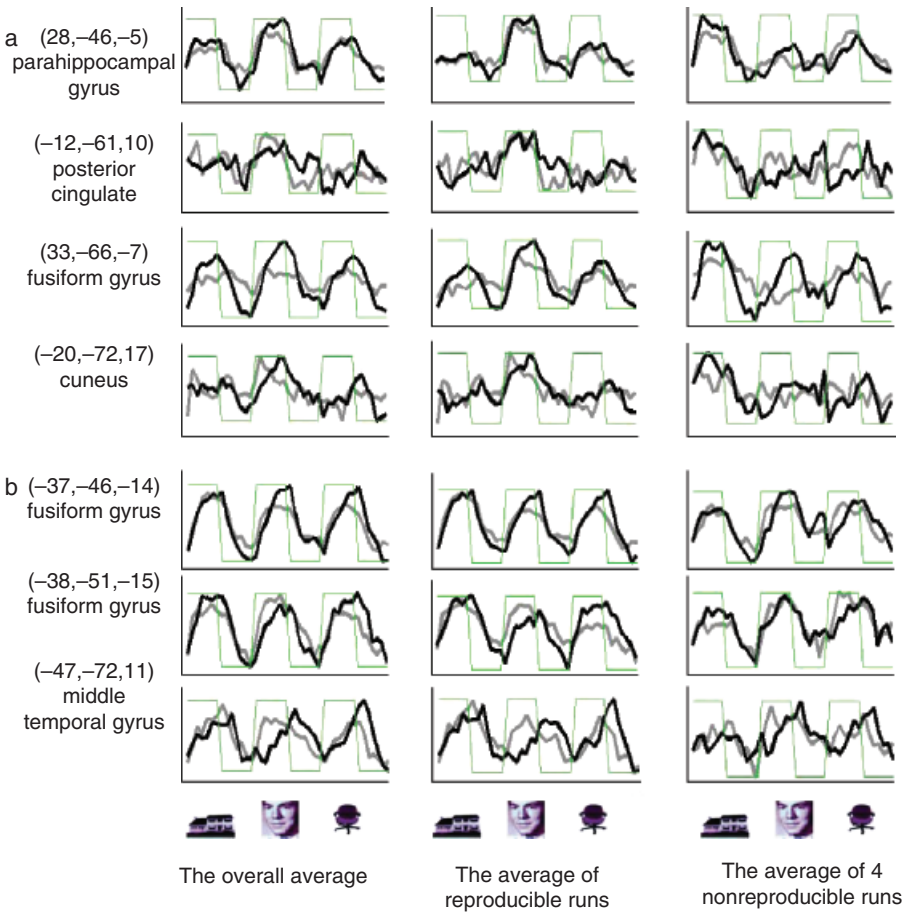

Figure 7. The HRFs of category-preferential regions for Subject 6. The HRFs corresponding to different regions are the averages of images across runs according to the original design sequence: P3 → V1 → P5 → V4 → P1 → V6 → P4 → V2 → P6 → V3 → P2 → V5. The four non-reproducible runs in (a) refer to V1, V5, P2, and P4, and in (b), refer to V3, V5, P4, and P6.

attentional deficit under Parkinson's disease has problems in execution of neuronal cells and disruption of activation processes. Under this hypothesis, schizophrenia patients in matching and change detection tasks will engage increased responses only in cuneus and precuneus. However, patients with Parkinson's disease will engage decreased or no response in the two regions. The proposed applications remain to be verified in further studies using reproducibility analysis.

REFERENCES

Blakemore, C., & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research*, *15*, 439–440.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Constable, R. T., Skudlarski, P., & Gore, J. C. (1995). An ROC approach for evaluating functional brain MR imaging and postprocessing protocols. *Magnetic Resonance in Medicine*, *34*, 57–64.

de Gelder, B., & Rouw, W. (2001). Beyond localisation: A dynamical dual route account of face recognition. *Acta Psychologica*, *107*, 183–207.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, *16*, 465–483.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*, 191–197.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*, 870–878.

Genovese, C. R., Noll, D. C., & Eddy, W. F. (1997). Estimating test-retest reliability in functional MR imaging I: Statistical methodology. *Magnetic Resonance in Medicine*, *38*, 497–507.

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews: Neuroscience*, *2*, 685–694.

Harrison, B. J., Yücel, M., Pujol, J., & Pantelis, C. (2007). Task-induced deactivation of midline cortical regions in schizophrenia assessed with fMRI. *Schizophrenia Research*, *91*, 82–86.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534.

Ishai, A., Ungerleider, L. G., Martin, A., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, *12*S2, 35–51.

Ishai, A., Ungerleider, L. G., Martin, A., Shouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Science, USA*, *96*, 9379–9384.

Ivanitskii, A. M. (1996). The cerebral basis of subjective experiences: The hypothesis of information synthesis. *Zh Vyssh Nerv Deiat Im I P Pavlova*, *46*, 241–252.

Ivanitskii, A. M., Strelets, V. B., & Korsakov, I. A. (1984). Brain informational processes and mental activity. *Moscow: Science* (in Russian).

Joseph, J. E., & Gathers, A. D. (2002). Natural and manufactured objects activate the fusiform face area. *NeuroReport*, *13*, 935–938.

Keil, A., Moratti, S., Sabatinelli, D., Bradley, M. M., & Lang, P. J. (2005). Additive effects of emotional content and spatial selective attention on electrocortical facilitation. *Cerebral Cortex*, *15*, 1187–1197.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation and action. In P. J. Lang, R. F. Simons, & M. Balaban (Eds.), *Attention orienting: Sensory and motivational Processes* (pp. 97–135). Hillsdale, NJ: Lawrence Erlbaum.

Li, C. S. R., Yan, P., Bergquist, K. L., & Sinha, R. (2007). Greater activation of the default brain regions predicts stop signal errors. *NeuroImage*, *38*, 640–648.

Liou, M., Su, H. R., Lee, J. D., Aston, J. A. D., Tsai, A. C., & Cheng, P. E. (2006). A method for generating reproducibility evidence in fMRI studies. *NeuroImage*, *29*, 383–395.

Liou, M., Su, H. R., Lee, J. D., Cheng, P. E., Huang, C. C., & Tsai, A. C. (2003). Bridging functional MR images and scientific inference: Reproducibility maps. *Journal of Cognitive Neuroscience*, *15*, 935–945.

Lubow, R. E. (1989). *Latent inhibition and conditioned attention theory*. NY: Cambridge University Press.

Luria, A. R. (1964). Factors and forms of aphasia. *CIBA Foundation Symposium on Disorders of Language*. London.

Mangina, C. A., & Sokolov, E. N. (2006). Neuronal plasticity in memory and learning abilities: Theoretical position and selective review. *International Journal of Psychophysiology*, *60*, 203–214.

Mechelli, A., Gorno-Tempini, M. L., & Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: Consistencies, inconsistencies and limitations. *Journal of Cognitive Neuroscience*, *15*, 260–271.

Olson, C. R., Musil, S. Y., & Goldberg, M. E. (1996). Single neurons in posterior cingulate cortex of behaving macaque: Eye movement signals. *Journal of Neurophysiology*, *76*, 3285–3300.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences, USA*, *98*, 676–682.

Saha, S., Long, C. J., Brown, E., Aminoff, E., Bar, M., & Solo, V. (2004). Hemodynamic transfer function estimation with Laguerre polynomials and confidence intervals construction from functional magnetic resonance imaging (fMRI) data. *IEEE ICASSP*, *3*, 109–112.

Scott, A. H., Guzeldere, G., & McCarthy, G. (2001). Dissociating neural mechanisms of visual attention in change detection using functional MRI. *Journal of Cognitive Neuroscience*, *13*, 1006–1018.

Shulman, A., Yacoub, E., Pfeuffer, J., van de Moortele, P. F., Adriany, G., Hu, X., & Ugurbil, K. (2002). Sustained negative BOLD, blood flow and oxygen consumption response and its coupling to the positive response in the human brain. *Neuron*, *36*, 1195–1210.

Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). ROC analysis of statistical methods used in functional MRI: Individual subjects. *NeuroImage*, *9*, 311–329.

Sokolov, E. N., Nezlina, N. I., Polyanskii, V. B., & Evtikhin, D. V. (2002). The orientating reflex: The targeting reaction and searchlight of attention. *Neuroscience Behavioral Physiology*, *32*, 347–362.

Strother, S., La Conte, S., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S., & Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage*, *23*S1, 196–207.

Su, H. R., Liou, M., Cheng, P. E., Aston, J. A. D., & Lai, S. H. (2007). Reproducibility analysis of event-related fMRI experiments using Laguerre polynomials. In M. Ishikawa, K. Doya, H. Miyamoto, & T. Yamakawa (Eds). *Lecture Notes in Computer Science: Neural Information Processing*. New York: Springer.

Swallow, K. M., Braver, T. S., Snyder, A. Z., Speer, N. K., & Zacks, J. M. (2003). Reliability of functional localization using fMRI. *NeuroImage*, *20*, 1561–1577.

Tulving, E., Markowitsch, H. J., Kapur, S., Habib, R., & Houle, S. (1994). Novelty encoding networks in the human brain: Positron emission tomography data. *NeuroReport*, *5*, 2525–2528.

Vygotsky, L. S. (1931). Psychology and conception of localization of psychical functions. *Moscow: Science* (In Russian).

Wojciulik, E., Kanwisher, N., & Driver, J. (1998). Covert visual attention modulates face-specific activity in the human fusiform gyrus: fMRI study. *Journal of Neurophysiology*, *79*, 1574–1578.

Worsley, K. J., Liao, C., Aston, J. A. D., Petre, V., Duncan, G., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, *15*, 1–15.