Bridging Functional MR Images and Scientific Inference: Reproducibility Maps

Michelle Liou[1,2], Hong-Ren Su[1], Juin-Der Lee[1], Philip E. Cheng[1], Chien-Chih Huang[1], and Chih-Hsin Tsai[1]

Abstract

■ Historically, reproducibility has been the sine qua non of experimental findings that are considered to be scientifically useful. Typically, findings from functional magnetic resonance imaging (fMRI) studies are assessed with statistical parametric maps (SPMs) using a p value threshold. However, a smaller p value does not imply that the observed result will be reproducible. In this study, we suggest interpreting SPMs in conjunction with reproducibility evidence. Reproducibility is defined as the extent to which the active status of a voxel remains the same across replicates conducted under the same conditions. We propose a methodology for assessing reproducibility in functional MR images without conducting separate experiments. Our procedures include the empirical Bayes method for estimating effects due to experimental stimuli, the threshold optimization procedure for assigning voxels to the active status, and the construction of reproducibility maps. In an empirical example, we implemented the proposed methodology to construct reproducibility maps based on data from the study by Ishai et al. (2000). The original experiments involved 12 human subjects and investigated brain regions most responsive to visual presentation of 3 categories of objects: faces, houses, and chairs. The brain regions identified included occipital, temporal, and fusiform gyri. Using our reproducibility analysis, we found that subjects in one of the experiments exercised at least 2 mechanisms in responding to visual objects when performing alternately matching and passive tasks. One gave activation maps closer to those reported in Ishai et al., and the other had related regions in the precuneus and posterior cingulate. The patterns of activated regions are reproducible for at least 4 out of 6 subjects involved in the experiment. Empirical application of the proposed methodology suggests that human brains exhibit different strategies to accomplish experimental tasks when responding to stimuli. It is important to correlate activations to subjects' behavior such as reaction time and response accuracy. Also, the latency between the stimulus presentation and the peak of the hemodynamic response function varies considerably among individual subjects according to types of stimuli and experimental tasks. These variations per se also deserve scientific inquiries. We conclude by discussing research directions relevant to reproducibility evidence in fMRI. ■

INTRODUCTION

Functional magnetic resonance imaging (fMRI) has emerged as an important technique for research into human brain function. Many fMRI studies use an on-and-off paradigm in which subjects respond alternately to experimental and control stimuli. Usually, a statistical method is applied at each voxel in space to test for significant differences between the on and off conditions. Findings from fMRI experiments are typically assessed with the statistical parametric maps (SPMs), which are reported by showing anatomy in the background, with colored overlays indicating those voxels with a level of significance exceeding a p value threshold (e.g., $p < .05$). Those suprathresholded voxels are, presumably, brain regions that are most responsive to the experimental stimuli. SPM-based inference with a p value threshold makes an assumption that localized changes in signal intensity differ as the subject performs experimental and control tasks. This assumption can be easily violated due to a variety of effects. In practice, error can be separated from experimental effects through properly designed replications and the use of valid statistical methods.

The sources of error in functional MR data have been discussed widely in the literature (e.g., Savoy, 2001; Genovese, Noll, & Eddy, 1997). In general, errors confounding experimental effects can be grouped into three types. First, transient errors affect individual responses in the same way within a given session, but they vary across distinct experimental sessions (Thye, 2000; Genovese et al., 1997). These errors are caused by idiosyncratic disturbances in the environment and add between-session variations to functional images (Skudlarski, Constable, & Gore, 1999). For example, subjects

[1]Academia Sinica, Taipei, Taiwan, [2]Fu-Jen Catholic University, Taipei, Taiwan

may become less attentive to stimuli due to fatigue or drowsiness; stimulus sequence may have unexpected order effects upon responses; or functional images may be impaired by global changes of intensity between sessions. Second, environmental, physiological, and psychological factors randomly fluctuate throughout the course of an experiment on a moment-by-moment basis. The occurrence of those random effects can be assumed to be equally likely across experimental sessions. Third, factors related to imaging techniques also affect the quality of observations. These factors include pulse sequence, imaging parameters, and scanner performance. Because errors are less likely to be reproducible, it is possible to assess these errors through analyses of reproducibility. Errors due to imaging techniques, on the other hand, can also be minimized through proper acquisition scheme selection. In this article, we particularly make a distinction between transient and random errors and will show that an estimate of between-session errors can reduce bias in estimating experimental effects in individual sessions.

Reproducibility, historically, has been the sine qua non of experimental findings that are considered to be scientific (Nickerson, 2000; Smith, Best, Cylke, & Stubbs, 2000; Branch, 1999; Carver, 1993). In fMRI studies, reproducibility requires that the same local activation maps are likely to be observed in an experimental replication. There is a common belief that a smaller p value implies a stronger likelihood of getting the same results on another replication of the same experiment. Carver (1978) referred to this belief as the ''reproducibility fantasy'' and contended that ''nothing in the logic of statistics allows a statistically significant result to be interpreted as directly reflecting the probability that the result can be replicated.'' Clearly, a smaller p value does not represent the complement of the likelihood that a result will be reproducible; all statistically significant findings must also provide evidence of reproducibility regarding the experimental outcome. Casey et al. (1998) studied reproducibility of fMRI results across four institutes using a spatial working memory tasks. Other studies evaluated analysis methods on reproducibility of the same SPMs (Salli, Korvenoja, Visa, Katila, & Aronen, 2001; Genovese et al., 1997; Noll et al., 1997). In this article, we focus on a methodology for assessing reproducibility without conducting separate experiments; we suggest that one should interpret SPMs in conjunction with evidence of reproducibility in fMRI studies.

fMRI experiments are usually performed over a period of time and are divided into smaller experimental sessions (or experimental runs) in order to allow subjects to rest. Image data are pooled across sessions and multiple subjects to construct the final SPMs (Skudlarski et al., 1999; Constable and Skudlarski, 1995). The classic paradox raised by Meehl (1967) remarked that the apparent power due to larger sample sizes also increases the possibility of making a Type I error. In fMRI studies with multiple sessions, reproducibility is easily assessed by appropriate information integration. The reproducibility of a voxel is defined here as the degree (number of times) to which the active status of the voxel, in responding to stimuli, remains the same across replicates (out of M experimental sessions) implemented under the same conditions. In this study, we also categorize voxels according to reproducibility; a voxel is strongly reproducible if its active status remains the same in at least 90% of the sessions, moderately reproducible in 70–90% of the sessions, weakly reproducible in 50–70% of the sessions, and otherwise not reproducible.

Statistical methods for analyzing fMRI data have to be sensitive to small signal changes (typically $<1\%$) and robust to mild violation of distributional assumptions. The general linear model has been commonly used for analyzing fMRI data in experiments involving multiple types of stimuli (Friston et al., 1995). In this study, we suggest augmenting the model slightly, by assuming that the model parameters in individual sessions are random samples from a known distribution. The augmented model is a special case of the empirical Bayes methodology, which provides a way of borrowing information across sessions to improve parameter estimates for each individual session (Rubin, 1980). In fMRI studies, experimental sessions may involve different tasks. The augmented model also allows for examining task effects on subjects' responses. The general linear model or the empirical Bayes method always generates voxelwise statistics (e.g., t values); individual voxels are assigned to the active/inactive status according to a threshold on the statistics. In this study, we suggest selecting a threshold by maximizing the overall reproducibility of active/inactive outcomes for all voxels. It has been observed frequently that, even with the same scanner and experimental paradigm, subjects can vary in the degree of activation. Therefore, different thresholds are appropriate for different subjects (Genovese, Lazar, & Nichols, 2002). The proposed threshold optimization procedure proceeds on an individual subject basis. In brief, the proposed methodology includes (i) the empirical Bayes method for estimating effects due to experimental stimuli, (ii) a threshold optimization procedure for assigning voxels to the active status, and (iii) construction of reproducibility maps.

The proposed methodology was implemented using data from the study by Ishai, Ungerleider, Martin, and Haxby (2000). These experiments involved visual presentation of three categories of objects: faces, houses, and chairs. The study found that a majority of voxels, which were maximally responsive to one of these objects, responded significantly to other objects as well (Ishai, Ungerleider, Martin, Shouten, & Haxby, 1999; Ishai et al., 2000). Here, we reanalyzed the data from the perspective of reproducibility. The reproducibility

maps proved to be a useful supplement to the SPMs in the Ishai et al. study. In the Methods, we present the methodology for constructing reproducibility maps, including the empirical Bayes method, the threshold optimization procedure, and map construction. In the Results, we present the reproducibility maps for the fMRI data in the Ishai et al. study. Finally, we suggest research directions that rest on reproducibility evidence in fMRI studies.

RESULTS

Figure 1a and b shows the receiver–operator characteristic (ROC) curves comparing results from the general

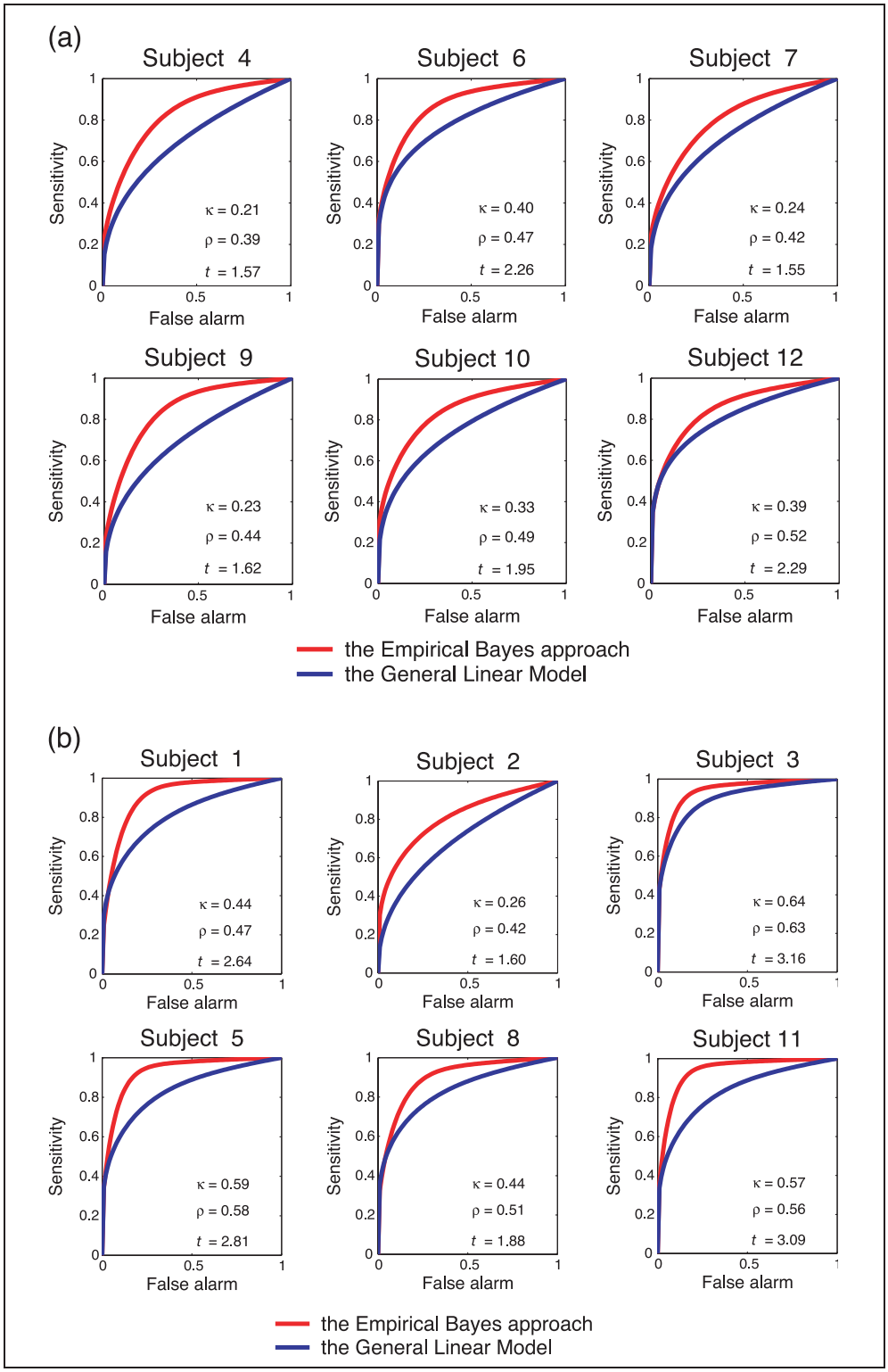

Figure 1. (a) ROC curves for the six subjects in Experiment 1. The ρ and κ indices and the threshold selected for each subject are also listed below the curves. (b) ROC curves and ρ and κ indices for the six subjects in Experiment 2.

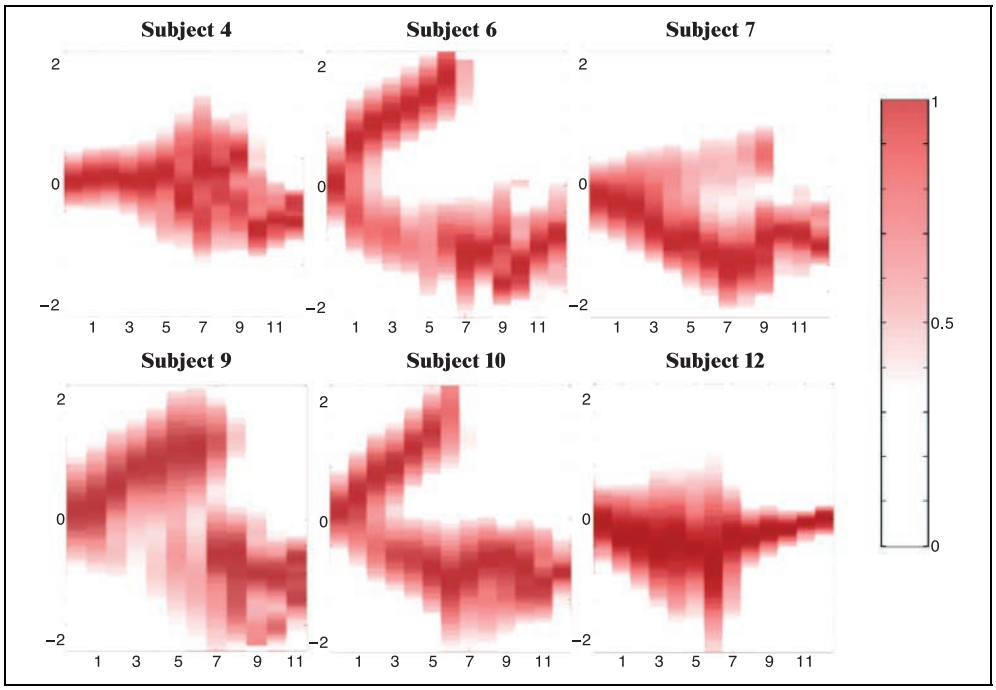

Figure 2. The distributions of t values comparing matching and passive tasks in Experiment 1. The distributions are grouped according to reproducibility of voxels. The t values are on the vertical axis and the reproducibility of voxels is on the horizontal axis.

linear model in Equation 4 and the empirical Bayes approach in assessing the objects versus control contrast. For the 12 subjects, the curves associated with the empirical Bayes approach all lie above those of the general linear model. In assessing reproducibility, the optimal decision has to be based on the β_i estimates in the individual session (i.e., computing R_v for each voxel). The ROC curves indicate that the empirical Bayes approach can improve sensitivity with little cost in terms of increasing the false alarm rate; the results are also free of thresholds. The ρ and κ indices are also listed in the figures for every subject. These indices were derived from results of the empirical Bayes analysis and are the maximum obtainable values given all possible thresholds. The κ values are generally smaller than the ρ values and closer to each other as κ becomes large (see the examples in Subjects 3, 5, and 11). It is also interesting to note that, on the average, reproducibility in Experiment 2 is stronger than in Experiment 1. In threshold optimization, a threshold for constructing reproducibility maps is a larger t value if the observed subject effects are reliable, and is a smaller t value otherwise (see a comparison between Subjects 2 and 3).

The voxel t values comparing effects between matching and passive tasks (i.e., effect sizes) in Experiment 1 are grouped according to voxels' reproducibility (i.e., the R_v values). The distributions of these t values according to reproducibility are plotted in Figure 2. The plots in the figure clearly show bimodal distributions for at least four out of six subjects. The patterns of these plots are robust to the threshold, that is, the same bimodal patterns persisted when the thresholds listed in Figure 1a were shifted to upper or lower bounds. These results indicate that there were at least two distinct mechanisms involved when subjects performed the matching and passive tasks. One mechanism has positively distributed t values and the other has negatively distributed t values. In the figure, Subjects 6 and 10 clearly show two mechanisms in performing the two tasks. The reproducibility maps for the two subjects are given in Figure 3. In constructing these maps, three-dimensional rendering was performed with the mri3dX software (http://www.aston.ac.uk/lhs/staff/singhkd/mri3dX/index). The colored voxels in these maps are either strongly or moderately reproducible. The maps suggest that a major portion of reproducible voxels are distributed in the temporal and occipital regions which concur with the SPMs in Ishai et al. (2000). The maps also show that the two subjects engage regions in the bilateral precuneus and posterior cingulate. Subject 6 also shows activations in the pre- and postcentral gyri and cerebellum. These regions all had a longer latency between the stimuli presentation and peak of the hemodynamic response function (HRF). The response waveforms for selected regions are plotted in Figure 4 for Subject 6. In delayed match-to-sample tasks, subjects indicated which choice stimulus matched a sample object by pressing a button with the right or left thumb. In passive viewing, subjects simply responded to stimuli without recording a sample stimulus or making a decision on choice stimuli. Ishai et al. (1999) indicated that the delayed matching task required more attention relative to passive viewing. According to the reproducibility maps in Figure 3, several regions engage in performing alternately the two tasks especially for Subject 6.

Figure 5 shows the densities of those t values comparing matching and drawing tasks in Experiment 2. The t values have unimodal distributions for all

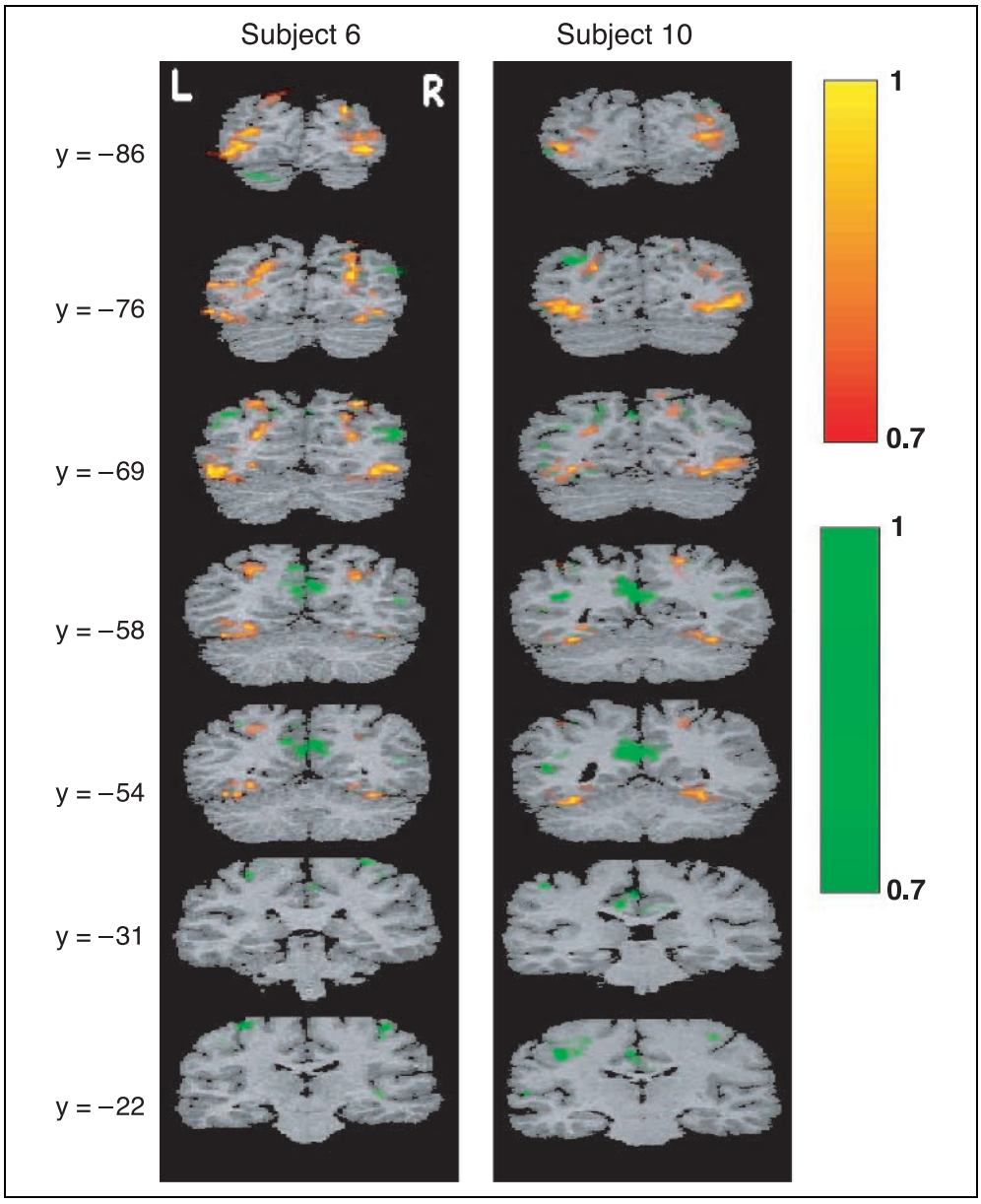

Figure 3. Reproducibility maps comparing Subjects 6 and 10 in Experiment 1. Coordinates are in the normalized space of the Talairach and Tournoux (1988) brain atlas. The selected slices are all in coronal sections. These *t* values plotted in Figure 2 are also referred to in this figure. Colored voxels in green regions have positively distributed *t* values and those in red regions have negatively distributed *t* values.

subjects. This result confirms the findings of Ishai et al. (2000) that photographs and line drawings evoked the same pattern of responses regardless of the low-level features of the stimuli, such as spatial frequency or texture. The *t* values for Subject 3 cluster around the positive side of the scale, which deviates slightly from other subjects in the same experiment. The reproducibility maps for Subjects 2 and 3 are shown in Figure 6. The region of activation for both subjects are also distributed in the temporal and occipital regions. Subject 2 had lower reproducibility values compared with other subjects. If we included weakly reproducible ($R_v/M > 50\%$) voxels, the active regions would have been even more similar. The waveforms of the hemodynamic response for selected regions are plotted in Figure 7. We note that data for Subject 3 gave the highest reliability values among all 12 subjects (see Figures 1 and 2). This subject also has a higher average HRF amplitude for drawing tasks compared with matching tasks.

DISCUSSION

Reproducibility evidence can be assessed in most fMRI studies conducted with multiple sessions. In this article, we have proposed a methodology for assessing the evidence with reproducibility maps in conjunction with SPMs. The proposed methodology includes (i) the empirical Bayes method, (ii) a threshold optimization procedure, and (iii) construction of reproducibility maps. We implemented the methods to reanalyze data from the study by Ishai et al. (1999, 2000) and derived reproducibility maps for those voxels that responded consistently to experimental stimuli. The maps were

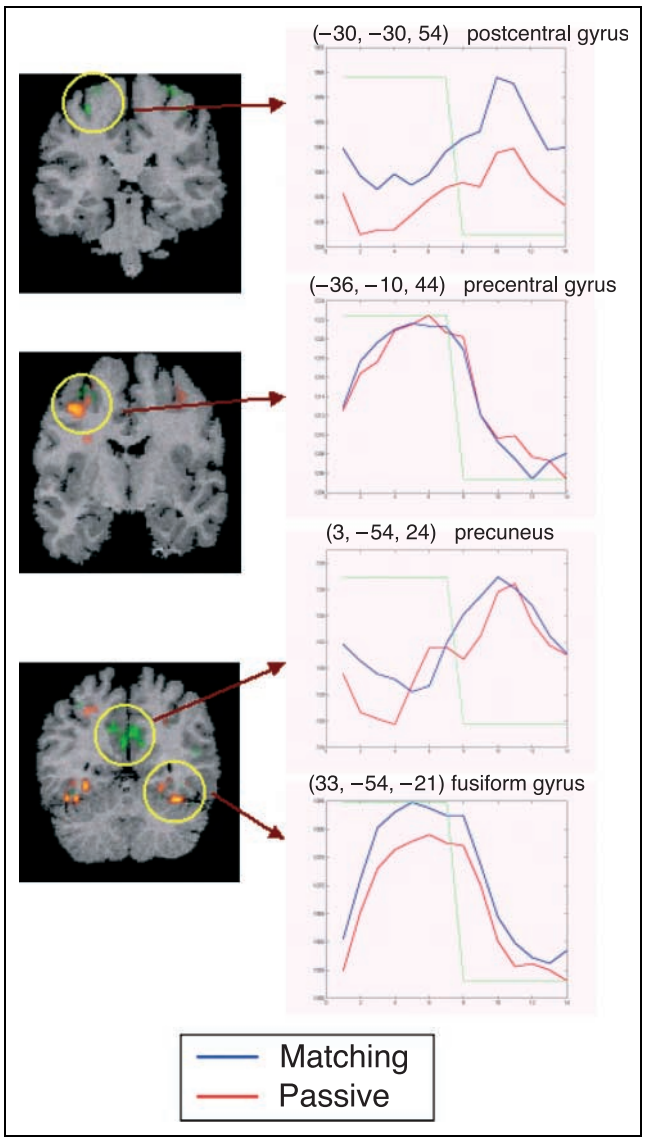

Figure 4. Forms of the hemodynamic response to objects in selected regions for Subject 6.

constructed for individual subjects without averaging data across subjects. The ensuing reproducibility maps suggest a few important findings. First, the amount of error differs in individual subjects. Experimental or statistical control over these errors may not make subjects completely interchangeable. It is important to consider separate thresholds for individual subjects, that is, less stringent thresholds for subjects that have larger errors. Reproducibility maps, constructed on an individual threshold basis, suggest experimental results that would have been observed if subjects were equally reliable. Second, subjects may exercise different strategies in performing experimental tasks while responding to stimuli. Based on reproducibility maps, brain regions evidencing task effects may be correlated with subjects' behavior such as reaction time and response accuracy. Furthermore, it is desirable to separate brain regions that are locked to experimental tasks from those locked to stimulus presentation. The proposed methodology finds different regions locked to various events and is generalizable to other more complicated designs. Third, even within the same brain region, latencies between stimulus presentation and the peak of the hemodynamic response vary considerably among subjects according to types of stimuli and experimental tasks. fMRI studies using the on-and-off paradigm have often incorporated an HRF model into the design matrix. The variations per se also deserve scientific inquiry. The use of the proposed methodology is not limited to the on-and-off designs. With event-related designs, for example, the fMRI data can be partitioned into smaller sessions, each being a replicate of the other; reproducibility can be assessed by an analogy of the proposed method. In conclusion, we have tried to take a step toward assessing reproducibility in fMRI studies without conducting further experiments, and we hope that research in this area will be encouraged.

METHODS

Experimental Data

In the empirical examples, fMRI data were collected from 12 subjects in two experiments (Ishai et al., 2000); the data sets were supported by the US National fMRI Data Center. The image data were preprocessed with correction for motion artifacts. These preprocessed data were analyzed in this study. The experiments examined the representation of objects in the human occipital and temporal cortex. In Experiment 1, six subjects were presented with gray-scale photographs of houses, faces, and chairs. For each subject, there were 12 experimental sessions that were subdivided into two tasks. In the passive viewing task, a stimulus (houses, faces, chairs) was presented at a rate of 2 sec followed by a phased, scrambled picture at the same rate, which served as the control stimulus. In the delayed matching task, a stimulus was followed, after a 0.5-sec delay, by a pair of choice stimuli presented at a rate of 2 sec. In Experiment 2, the other six subjects performed the delayed match-to-sample task with photographs and line drawings of houses, faces, and chairs. In the original reports (Ishai et al., 1999, 2000), three orthogonal contrasts were examined in the two experiments, namely, meaningful objects (i.e., faces, houses, and chairs) versus control stimuli (i.e., phased, scrambled pictures), faces versus houses/chairs, and houses versus chairs. In the current study, we only construct reproducibility maps for voxels that responded to all three objects (i.e., the objects versus control stimuli contrast).

The General Linear Model

In constructing SPMs with the general linear model, observations in each voxel are normally pooled over

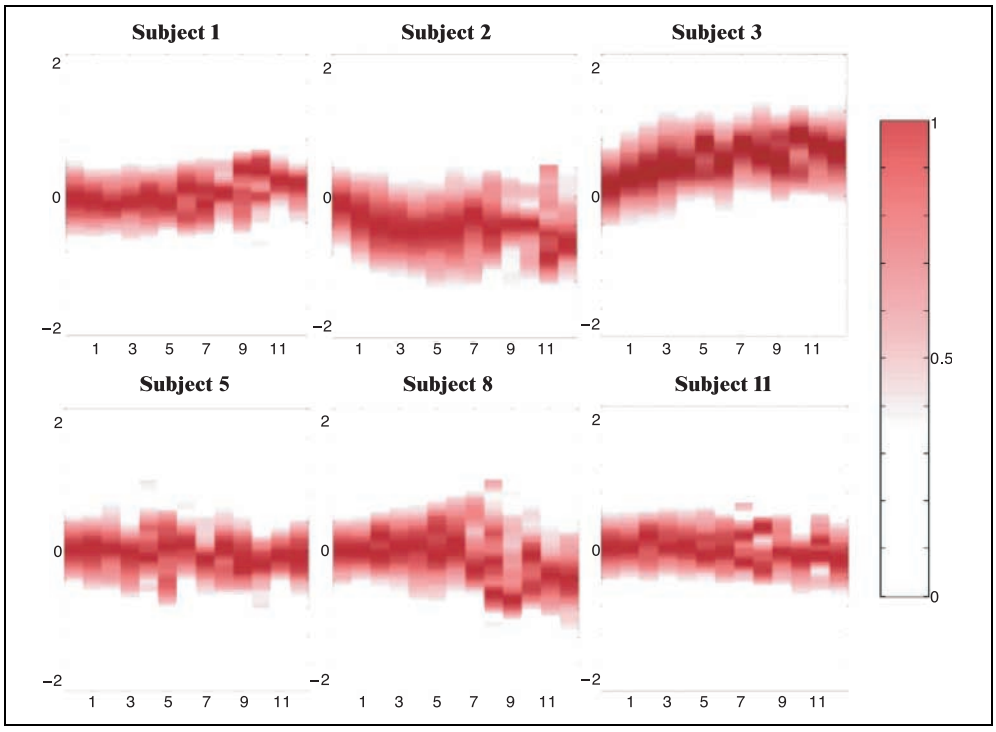

Figure 5. The distributions of t values comparing delayed matching with photographs and line drawings in Experiment 2. The t values are on the vertical axis and the reproducibility of voxels is on the horizontal axis. The ± values are on the vertical axis and the reproducibility of voxels is on the horizontal axis.

sessions before estimating model parameters. Let y be the vector of n pooled observations. The model assumes that

$$\mathbf{y} = \mathbf{X}\beta_{\text{pool}} + \mathrm{e} \tag{1}$$

where $\mathbf{X}$ is the design matrix and β_{pool} is the vector containing the unknown regression parameters. The design matrix contains indicator variables or regressors that correspond to the linear effects investigated in the fMRI experiment. Each contrast (a column in $\mathbf{X}$) has a corresponding parameter in the β_{pool} vector. In order to apply a statistical test, the y observations are assumed to have a Gaussian distribution with mean $\mathbf{X}\beta_{\text{pool}}$ and variance $\sigma^2\mathbf{I}_n$, where σ^2 is the residual variance and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. The standard least-squares method gives the following β_{pool} and σ^2 estimates:

$$\hat{\beta}_{\text{pool}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n-k} \| y - \mathbf{X}\hat{\beta}_{\text{pool}} \|^2 \tag{3}$$

where k is the number of columns in $\mathbf{X}$ and $\mathbf{X}'$ denotes the transposition of $\mathbf{X}$. Inferences about each parameter can be made by computing a t statistic, a ratio that compares the parameter estimate with the standard error of the estimate. The standard error of each $\hat{\beta}_{\text{pool}}$ estimate is the corresponding diagonal element in $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, which is the variance of $\hat{\beta}_{\text{pool}}$. The associated p value is the probability of exceeding an observed t value in a Student's t distribution with $n - k$ degrees of freedom. Each voxel is classified active/inactive according to a preselected p value (e.g., $p < 10^{-4}$ in the study of Ishai et al., 2000). This p value can be adjusted for multiple voxels using the Random Field Theory as discussed by Friston et al. (1994; 1995).

In fMRI experiments, $\hat{\beta}_{\text{pool}}$ is estimated under the assumption that observations in individual sessions are interchangeable. Empirical studies have suggested averaging the images from M experimental sessions and then calculating t statistics using those averaged images. If experimental sessions differ in order of stimulus presentation (order of faces, houses, and chairs), as well as task performance (matching vs. drawing), individual regression parameters should be obtained from each session separately (Skudlarski et al., 1999; Constable and Skudlarski, 1995). This procedure is effective because individual statistics are not affected by substantial variations among sessions. Specifically, Equation 1 can be modified for each session as

$$y_i = \mathbf{X}_i\beta_i + \mathrm{e}_i \tag{4}$$

where $\mathbf{X}_i$ denotes the partition on submatrix in $\mathbf{X}$ corresponding to the ith session. Here, we still assume that y_i has a Gaussian distribution with mean $\mathbf{X}_i\beta_i$ and variance $\sigma_i{}^2\mathbf{I}_{n_i}$. Researchers have suggested using t statistics based on Equation 4 to compare test–retest reliability in functional MR images across sessions (Maitra, Roys, & Gullapilli, 2002; Skudlarski et al., 1999; Genovese et al., 1997).

The Empirical Bayes Method

It is generally desirable to assign different weights to individual β_i's in Equation 4 for combining these estimates to assess the responses over sessions. Here, we

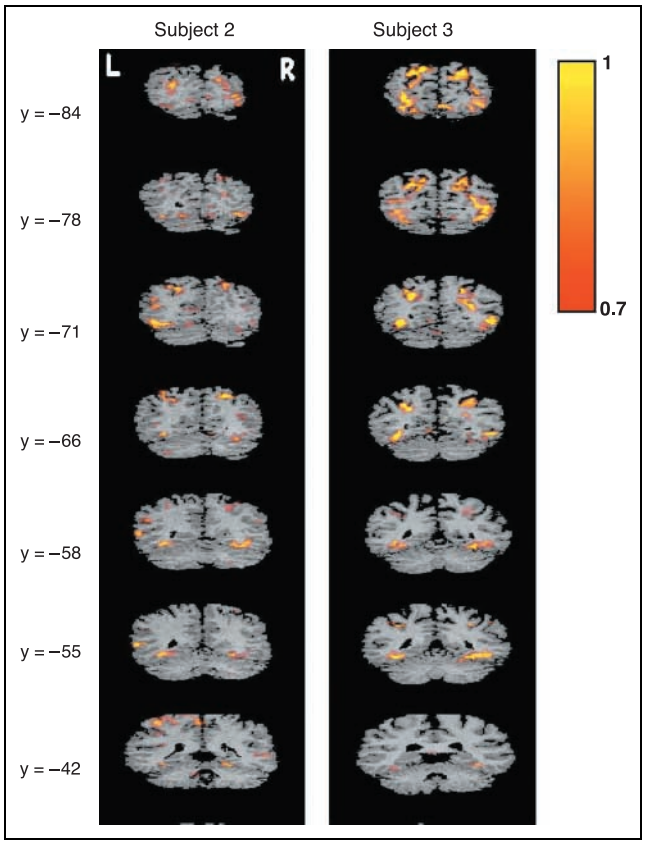

Figure 6. Reproducibility maps comparing Subjects 2 and 3 in Experiment 2. Voxels in colored regions are at least moderately reproducible according to the definition in the main text.

propose the empirical Bayes method for weighing information across sessions. The usual empirical Bayes estimate of β_i can be represented as a weighted combination of contributions from sessions as a whole (i.e., β_{pool}) and the individual session (i.e., β_i). Essentially, the method provides a way for borrowing information across sessions to rectify biased estimates at each individual session (Rubin, 1980). With the method, Equation 4 is augmented by assuming a priori the β_i for $i = 1, \ldots, M$ ($M = 12$ in the work of Ishai et al., 2000) are random samples from a multivariate Gaussian distribution with mean μ_i and a common variance–covariance matrix Ω.

In fMRI applications, different experimental sessions may involve separate tasks. For example, subjects in Experiment 1 performed six delayed match-to-sample tasks and six passive viewing tasks. It would also be interesting to estimate β_i according to the types of tasks. Let $\mathbf{B}$ be a $k^* \times k$ matrix containing the multivariate regression parameters, where k^* denotes the number of task effects examined between sessions and k is the length of β_i. We assume that

$$\beta_i = \mathbf{B}' x_i^* + \nu_i \tag{5}$$

where the transposition of x_i^* is the ith row in $\mathbf{X}^*$, which is the design matrix for estimating β_i. (Note that $\mathbf{X}_i$ in Equation 4 is the design matrix for estimating the waveform in the ith voxel, and $\mathbf{X}^*$ is the design matrix for estimating β_i for $i = 1, \ldots, M$.) Based on the model, the mean of β_i is $\mu_i = \mathbf{B}' x_i^*$ and the variance of ν_i is the common Ω. Given our assumptions, the posterior expectation of β_i can be expressed as:

$$E(\beta_i \mid \sigma_i^2, \mu_i, \Omega, \mathbf{X}_i, y_i) \equiv h_i$$

$$= \left(\Omega^{-1} + \sigma_i^{-2}(\mathbf{X}_i'\mathbf{X}_i)\right)^{-1}\left(\Omega^{-1}\mu_i + \sigma_i^{-2}\mathbf{X}_i' y_i\right) \tag{6}$$

and its cross product (i.e., the conditional covariance) is

$$E(\beta_i \beta_i' \mid \sigma_i^2, \mu, \Omega, \mathbf{X}_i, y_i) \equiv H_i$$

$$= \left(\Omega^{-1} + \sigma_i^{-2}(\mathbf{X}_i'\mathbf{X}_i)\right)^{-1} + h_i h_i' \tag{7}$$

The estimation of σ_i^2, μ_i, and Ω can be accomplished via iterative Expectation Maximization (EM). Equations 6 and 7 constitute the E-step in the algorithm. We briefly outline the M-step for estimating unknown parameters. Let $\mathbf{Z}$ be a matrix containing all estimated β_i in Equation 6, that is, $\mathbf{Z}' = [h_1, h_2, \ldots, h_M]$. At the $(r + 1)$th iteration, the algorithm computes

$$\mu_i^{(r+1)} = \mathbf{B}'^{(r)} x_i^*, \text{where } \mathbf{B}^{(r)} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Z}^{(r)} \tag{8}$$

$$\Omega^{(r+1)} = \frac{1}{M - k^*}\left\{\sum_i H_i^{(r)} - \frac{1}{M}\mu^{(r)}\mu'^{(r)}\right\}, \text{where}$$

$$\mu^{(r)} = \sum_i \mu_i^{(r)}, \text{and} \tag{9}$$

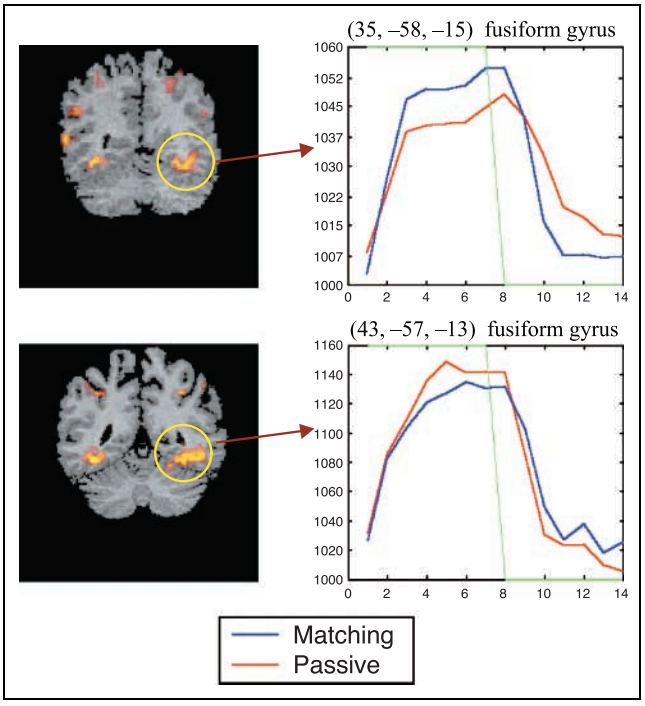

Figure 7. Forms of the hemodynamic response to objects in selected regions for Subject 2 (top) and Subject 3 (bottom).

$$\sigma_i^{2(r+1)} = \frac{1}{n-k} \| y_i - \mathbf{X}_i \beta_i^{(r)} \|^2 \qquad (10)$$

The E- and M-steps are iterated until the sequence of parameter estimates converges. An interested reader may refer to Rubin (1980) and Dempster, Laird, and Rubin (1977) for the rationale underlying EM procedures. Friston et al. (2002) also detailed the EM procedures for Bayesian inference in neuroimaging and suggested using a weighted least squared estimate for $\mathbf{B}^{(r)}$ in Equation 8. Because the data are balanced such that each session has the same number of observations, the two estimates will not differ dramatically.

By analogy with the general linear model, the t statistics for empirical Bayes estimates of the β_i in Equation 6 can be computed by the use of the corresponding posterior standard deviations given the maximum likelihood estimates of μ_i and Ω. The analyses below are restricted to those t values that test responses to all three stimuli. (Note: For each voxel, there are 12 t values, 1 for each session.) Alternatively, t statistics can also be computed for parameters in $\mathbf{B}$. Here, we are interested in the effect due to different tasks, that is, the indicator variable in $\mathbf{X}^*$ with score ''1'' for the passive task and ''−1'' for the matching task in Experiment 1, and score ''1'' for the drawing task and ''−1'' for the matching task in Experiment 2. The t values correspond to ''random-effects'' analysis because they test the effect size against the session-to-session variability Ω that is treated as a random effect. (Note: For each voxel, there is one t value corresponding to the task effect.)

ROC Analyses of Statistical Methods

Statistical methods can be evaluated in their capacity to differentiate the truly active voxels from the truly inactive voxels. The ROC approach has been recommended for comparing statistical methods in fMRI data analyses (Maitra et al., 2002; Skudlarski et al., 1999; Genovese et al., 1997; Friston, Holmes, Poline, Price, & Frith, 1996). Sensitivity is defined as the proportion of truly active voxels that are classified active. This proportion represents the power of a statistical method. False alarm rate, on the other hand, is the proportion of truly inactive voxels that are classified active and contribute to Type I error. Sensitivity can always be increased by lowering the threshold, a situation in which the false alarm rate is inflated. In fMRI studies, the true status of each voxel is unknown, but the two proportions can be estimated from the data. Genovese et al. (1997) suggested estimating the proportions at a particular threshold by assuming a mixed binomial model underlying the number of times (out of M replications) that a voxel is consistently classified active. Let p_A and p_I denote sensitivity and false alarm rate, respectively, and R_v represents the number of replications, out of M, that a voxel is classified active (R_v refers to the reproducibility of the vth voxel by definition). The mixed binomial model assumes that R_v is a random sample from

$$\binom{M}{R_v} \left[\lambda p_A^{R_v} (1-p_A)^{(M-R_v)} + (1-\lambda) p_I^{R_v} (1-p_I)^{(M-R_v)} \right] \qquad (11)$$

where λ is the proportion of truly active voxels. The model in Equation 11 can easily be generalized to multinomial cases in which there is one λ parameter and several threshold values, each with its corresponding paired (p_A, p_I) parameters.

The ROC curve is a bivariate plot of sensitivity against false alarm rate for different thresholds. The sensitivity of a particular statistical method can also vary according to subjects, stimulus conditions and other unknown factors. Maitra et al. (2002) extended the model in Equation 11 by incorporating spatial dependence among nearby voxels. The extended model gave slightly more conservative estimates of p_A and p_I. Our analyses used a maximum likelihood procedure to estimate the λ and paired (p_A, p_I) parameters for selected thresholds; this mixed multinomial model is similar to the proposal of Genovese et al. (1997, p. 507). We also used the exponential model suggested by England (1988) for smoothing and extrapolating the ROC curves that were interpreted in a relative fashion between methods and subjects in the empirical example.

Decision Threshold Optimization

In fMRI experiments, it is important to understand the extent to which replicates made under the same conditions give the same results. The observed proportion of agreement between the true active/inactive status and classification result is $p_O = \lambda p_A + (1-\lambda)(1-p_I)$ This proportion can be corrected for chance, which is found by summing over the agreement diagonals, the product of the proportions for the row and column of the cell. We denote the agreement expected by chance as $p_C = \lambda\tau + (1-\lambda)(1-\tau)$, where $\tau = \lambda p_A + (1-\lambda)p_I$. Using the ROC model, the proportion of agreement corrected for chance is

$$\rho = \frac{p_O - p_C}{1 - p_C} \qquad (12)$$

which is the Kappa index due to Cohen (1960). The threshold value at the maximum ρ can be selected as the optimal operating point on the ROC curve. For a total of k contrasts in the design matrix, there will be 2^k possible

outcomes with combinations of the active/inactive status. With the ROC approach, the optimal threshold must be selected for each contrast. An alternative approach would identify the optimal operating points at maximum reproducibility in the 2^k outcomes across the M replications. With this approach, the k thresholds can be identified simultaneously without use of the ROC models. In the literature, there are several indices useful for assessing the reproducibility of categorical outcomes, for example, the Kappa-type coefficients (Roberts & McNamee, 1998; Posner, Sampson, Caplan, Ward, & Cheney, 1990; Shouten, 1980; Fleiss & Cohen, 1973) Let n_{ij} be the number of times that the ith voxel is assigned to the jth outcome out of M replications. Following Shouten (1980), the agreement between outcomes j and k can be assessed by

$$\kappa_{jk} = \frac{P_j P_k - \sum_i n_{ij} n_{ik} / [M(M-1)V]}{P_j P_k} \qquad (13)$$

where P_j is the sum of $n_{ij}/(MV)$ across V voxels and P_k is computed by analogy. The reproducibility of outcomes can be summarized by a weighted average of κ_{jk}, that is,

$$\kappa = \frac{\sum_{j \neq k} P_j P_k \kappa_{jk}}{\sum_{j \neq k} P_j P_k} \qquad (14)$$

The optimal thresholds for the k contrasts can be decided by maximizing κ. In empirical applications, we have found that the solution to λ in the mixed multinomial model is not unique, but in a small range. These λ estimates gave almost identical ROC curves. In the empirical example below, the optimal threshold was selected based on κ, but values of both sorts of indices are presented. The maximum ρ value was computed using the average of λ estimates.

Reproducibility Maps

Given an optimal threshold, the truly active voxels must be strongly reproducible (i.e., R_v/M above 90%). Some truly active voxels may exhibit moderate reproducibility (i.e., R_v/M between 70% and 90%) due to errors in estimating t values. In our empirical examples, we selected strongly reproducible voxels to construct the reproducibility maps but included voxels that were moderately reproducible and spatially proximate to strongly reproducible voxels.

Acknowledgments

The authors are indebted to Professor Karl J. Friston and anonymous reviewers for valuable comments on an earlier version of this manuscript. This research was supported by grant NSC89-2413-H-001-007 from the National Science Council, Taiwan.

Reprint requests should be sent to Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, or via e-mail: mliou@stat.sinica.edu.tw.

REFERENCES

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22,* 87–92.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287–292.

Casey, B. J., Cohen, J. D., O'Craven, K., Davidson, R. J., Irwin, W., Nelson, C. A., Noll, D. C., Hu, X., Lowe, M. J., Rosen, B. R., Truwitt, C. L., & Turski, P. A. (1998). Reproducibility of fMRI results across four institutes using a spatial working memory task. *Neuroimage, 8,* 249–261.

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

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *The Journal of the Royal Statistical Society (B), 39,* 1–38.

England, W. L. (1988). An exponential model used for optimal threshold selection on ROC curves. *Medical Decision Making, 8,* 120–131.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 20,* 37–46.

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *Neuroimage, 4,* 223–235.

Friston, K. J., Holmes, A., Worsley, K. J., Poline, J. B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2,* 189–210.

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.

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C., (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping, 1,* 241–220.

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.

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,* 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 Sciences, U.S.A., 96,* 9379–9384.
Maitra, R., Roys, S. R., & Gullapalli, R. P. (2002). Test–retest reliability estimation of functional MRI data. *Magnetic Resonance in Medicine, 48,* 62–70.
Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.
Noll, D. C., Genovese, C. R., Nystrom, L. E., Vazquez, A. L., Forman, S. D., Eddy, W. F., & Cohen, J. D. (1997). Estimating test–retest reliability in functional MR imaging: II. Application to motor and cognitive activation studies. *Magnetic Resonance in Medicine, 38,* 508–517.
Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., & Cheney, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine, 9,* 1103–1115.
Roberts, C., & McNamee, R. (1998). A matrix of kappa-type coefficients to assess the reliability of nominal scales. *Statistics in Medicine, 17,* 471–488.
Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association, 75,* 801–827.
Salli, E., Korvenoja, A., Visa, A., Katila, T., & Aronen, H. J. (2001). Reproducibility of fMRI: Effect of the use of contextual information. *Neuroimage 13,* 459–471.
Savoy, R. L. (2001). History and future directions of human brain mapping and functional neuroimaging. *Acta Psychologica, 107,* 9–42.
Shouten, H. J. A. (1980). Nominal scale agreement among observers. *Psychometrika, 51,* 453–466.
Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). ROC analysis of statistical methods used in functional MRI: Individual subjects. *Neuroimage, 9,* 311–329.
Smith, L. D., Best, L. A., Cylke, V. A., & Stubbs, D. A. (2000). Psychology without *p* values. *American Psychologist, 55,* 260–263.
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxis atlas of the human brain* (M. Rayport, Trans.). New York: Thieme.
Thye, S. R. (2000). Reliability in experimental sociology. *Social Forces, 78,* 1277–1309.