Cogn Affect Behav Neurosci
DOI 10.3758/s13415-013-0195-1

fMRI reliability: Influences of task and experimental design

Craig M. Bennett · Michael B. Miller

Abstract As scientists, it is imperative that we understand not only the power of our research tools to yield results, but also their ability to obtain similar results over time. This study is an investigation into how common decisions made during the design and analysis of a functional magnetic resonance imaging (fMRI) study can influence the reliability of the statistical results. To that end, we gathered back-to-back test–retest fMRI data during an experiment involving multiple cognitive tasks (episodic recognition and two-back working memory) and multiple fMRI experimental designs (block, event-related genetic sequence, and event-related m-sequence). Using these data, we were able to investigate the relative influences of task, design, statistical contrast (task vs. rest, target vs. nontarget), and statistical thresholding (unthresholded, thresholded) on fMRI reliability, as measured by the intraclass correlation (ICC) coefficient. We also utilized data from a second study to investigate test–retest reliability after an extended, six-month interval. We found that all of the factors above were statistically significant, but that they had varying levels of influence on the observed ICC values. We also found that these factors could interact, increasing or decreasing the relative reliability of certain Task × Design combinations. The results suggest that fMRI reliability is a complex construct whose value may be increased or decreased by specific combinations of factors.

Keywords fMRI statistics · Reliability

Electronic supplementary material The online version of this article (doi:10.3758/s13415-013-0195-1) contains supplementary material, that is available to authorized users.

C. M. Bennett (✉) · M. B. Miller
Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106, USA
e-mail: bennett@psych.ucsb.edu

Published online: 10 August 2013

Reliable results are of obvious utility in even the most basic of neuroimaging studies. After all, the ability to obtain consistent values across time represents a powerful statement regarding the observed effects in a study. However, quantifying reliability in neuroimaging data is notoriously difficult. In a very technical sense, each experimental combination of task, design, scanner, image-processing pipeline, and analysis strategy has a reliability value unique unto itself, one that may or may not generalize to other studies (Braver, Cole, & Yarkoni, 2010). However, there is still merit in evaluating the distribution of reliability values observed across these many combinations. For example, knowing empirically that many studies can achieve moderate reliability levels, but few studies can obtain excellent reliability, is crucial to consider when evaluating the results of a future study. As functional neuroimaging is increasingly used across a diverse set of research topics, including clinical diagnosis and outcome prediction (Brodersen et al., 2012; Matthews, Honey, & Bullmore, 2006; Sato, Hoexter, Fujita, & Rohde, 2012), a thorough understanding of the many factors that influence the reliability of between-subjects fMRI results is an absolute necessity.

To be clear, *reliability* is a quantitative measurement that indexes the stability of data values. This is typically calculated as stability over time for test–retest data, or stability of agreement across raters. An array of measures such as intraclass correlation (ICC), coefficient of variation, Cohen's kappa index, Kendall's *W*, Pearson correlation, and numerous others, have all been used to index the reliability of fMRI values over time. In contrast, *reproducibility* is a qualitative measure of the ability to obtain similar results over time. For fMRI, this is often calculated in terms of the spatial location or total volume of activation observed in statistical results. Measures such as the dice coefficient and the Jaccard overlap have been used to gauge the similarity of activation patterns between two time points.

In the last four years, several new aspects of neuroimaging have come under the microscope with regard to their impact on reliability. For example, the influence of development has been

explored, with reliability values increasing between childhood and adolescence (Koolschijn, Schel, de Rooij, Rombouts, & Crone, 2011). Within- and between-scanner reliability has also been examined, with between-scanner results being roughly equivalent to those from a single scanner (Brown et al., 2011; Gradin et al., 2010). Furthermore, head positioning has been shown not to significantly influence the reproducibility of activation (Soltysik et al., 2011). In short, significant progress has been made regarding how specific design and analysis decisions can influence the reliability of a study's results.

One issue that remains unresolved is the lack of studies that have investigated multiple factors simultaneously. It is true that the reliability of a diverse array of tasks has been investigated using a broad array of methods—over one hundred published articles have investigated some aspect of fMRI reliability (Bennett & Miller, 2010). Most of these previous investigations have examined the reliability of results from a single cognitive task in isolation. This is informative, as it yields information on how reliable the results of specific tasks can be. However, of much greater utility would be the joint comparison of multiple tasks simultaneously. This would allow for the results of each task to be compared and contrasted against each other. It would also allow for the examination of potential interactions between experimental factors.

Only a handful of studies have attempted to compare reliability across multiple tasks within the same investigation. Yetkin, McAuliffe, Cox, and Haughton (1996) were among the first to directly contrast the reliability of multiple tasks. They examined the test–retest results from a sensory and a motor task in a group of four subjects, to find that the motor task had a higher proportion of jointly active pixels than did the sensory task. Later, Waldvogel, van Gelderen, Immisch, Pfeiffer, and Hallett (2000) examined the results of a visual and a motor task, to determine that intersession differences between the two tasks were correlated. Havel et al. (2006) examined a four-condition motor task to show that the results of different motor movements had varying levels of reliability. Harrington, Farias, Buonocore, and Yonelinas (2006) examined three types of memory-encoding tasks, to show that the reliability of results varied by encoding task and brain region. Caceras, Hall, Zelaya, Williams, and Mehta (2009) had the same set of subjects complete an auditory target detection task and an *N*-back task while in the scanner. Using their ICCmed index, Caceras et al. found that the *N*-back task had higher values both within the activated task network and across the whole brain. Several studies have examined the reliability of multiple language tasks. Harrington, Buonocore, and Farias (2006) examined six language tasks to determine the reliability of activation volume and laterality index. In their study, verb generation was associated with the highest concurrence ratio. Rau et al. (2007) used two language tasks to examine the reliability of activation within a specific region, Broca's area. They found that their task could not reliably identify Broca's area in their group of volunteers. Most recently, Raemaekers, du Plessis, Ramsey, Weusten, and Vink (2012) were able to compare the results from a visual and a motor inhibition task separated by a test–retest interval of one week. They used a novel method of estimating the BOLD signal variability between sessions to find that the differences in activation amplitude between sessions were 13.8 % for the visual task and 23.4 % for the motor inhibition task.

Although these previous studies have provided important information on the relative reliability of specific tasks, a number of issues remain with regard to interpreting their information. First, the methods used to quantify reliability in each study have varied dramatically. Some studies used intraclass correlations, some used Jaccard or dice overlap coefficients, and some were focused simply on the reliability of the language laterality index. Second, the studies varied widely with regard to their experimental designs. For example, one study had 18 subjects, whereas another had only four. Furthermore, some studies used event-related fMRI designs, whereas others used a block design. The variation in subject number and fMRI design means that the experiments had widely different values of statistical power from which to draw their conclusions. Statistical power, in this case, is the ability to detect relevant signals in a second-level group fMRI analysis. Although we know a great deal about how experimental design can impact statistical power (Liu, 2004; Liu & Frank, 2004), the relationship between power and reliability is less well defined. Existing evidence has shown only a modest relationship between the detection of legitimate results and the reliability of those results (Caceres et al., 2009; Zhang et al., 2009). In the context of these challenges, it is difficult to compare and contrast the relative influences of many experimental factors on fMRI reliability. Simultaneously manipulating several factors of interest within a single study could be very useful in quantifying their relative influences on the reliability of fMRI results. That was the goal of the present study.

The aim of this project was to examine the reliability of between-subjects differences in fMRI results under a series of manipulations involving the type of cognitive task, the fMRI experimental design, the contrast type, statistical thresholding, and the test–retest interval. To accomplish these goals, we examined reliability using data from two fMRI studies that each utilized a test–retest methodology. From these data, we were able to investigate the principal effects of these factors and their interactions on the reliability of the results.

Method

Primary study

Our primary data set was acquired to evaluate the influences of both cognitive task and fMRI experimental design on

estimates of reliability. This study included data from an episodic word recognition task and a two-back working memory task. Each task was completed using both a block and two event-related fMRI designs. This set up a 2 × 3 design, whereby we could simultaneously examine the contributions of task and design to the reliability of the results. The time between test and retest samples was approximately 20 min. Sixteen subjects participated in this study, with two subjects being excluded from the analysis for excessive head motion during acquisition (>1 mm per repetition time [TR]). Subjects' ages ranged from 19 to 35 years, with a mean age of 23.7. All subjects were right-handed and had normal or corrected-to-normal vision. They each completed an informed consent procedure that was approved by the UCSB Human Subjects Committee, and all were paid for their participation.

Task description

We utilized a standard episodic word recognition test and a two-back working memory test as the cognitive tasks for this experiment. Each task was considered to be a "target-versus-nontarget" task, with the subject remaining vigilant for the appearance of targets in a sequence of presented stimuli. The two tasks were designed to be as similar as possible, with each having two conditions of interest and equivalent numbers of stimuli per condition.

The episodic word recognition task was completed using separate encoding and recognition periods. The encoding of target words was completed within the scanner during anatomical image acquisitions. Each subject was shown a series of words and asked to try and remember them for later recognition. Immediately following the encoding period, the word recognition test would begin. During the recognition task, subjects were asked to report whether they believed each displayed word had been encountered during the encoding session (a target, or an "old" word), or whether they believed that the word had only been seen during the testing session (a nontarget, or a "new" word). Fifty old words and 50 new words were presented to the subjects, who responded using one of two buttons with their right hand. The presentation time for each stimulus was 1 s, and the task length was 5.5 min.

The *n*-back working memory task was presented as a two-back task involving a series of letters presented on the screen one at a time. Subjects were asked to report whether the currently displayed letter matched the letter that had been displayed two items previously in the sequence (a target) or whether it did not (a nontarget). Fifty targets and 50 nontargets were presented to each subject. Although this is a nonstandard proportion of targets to nontargets for a two-back task, it was necessary in order to make the stimulus counts equivalent between the two tasks. Subjects responded using one of two buttons with their right hand. The presentation time for each stimulus was 1 s, and the task length was 5.5 min.

fMRI experimental design

Each task was presented using three fMRI experimental designs. The first stimulus presentation condition utilized a block design. This design was chosen as the optimal approach to maximize statistical power and detection ability. The block design consisted of alternating 30-s epochs of task and rest. The frequency of stimulus presentation within each block maximized the degree of blood oxygenation level dependent (BOLD) signal shift in task-relevant regions of the brain, providing the greatest signal-to-noise ratio for identifying legitimate effects. Fifteen stimuli were presented at a rate of one every 2.0 s in each block. The block design length for this experiment was based on empirically derived optimized values for fMRI block designs (Maus, van Breukelen, Goebel, & Berger, 2010).

The second stimulus presentation condition was an event-related design generated by means of a genetic optimization algorithm. A genetic algorithm is able to search through the space of all possible stimulus timing combinations to find a subset of sequences that meet prespecified criteria. In this study, the genetic algorithm was weighted to provide a balance between statistical power and the ability to estimate the average hemodynamic response function for each condition. The results combined aspects of the block design (periods of rapid stimulus presentation) and an event-related m-sequence (low autocorrelation with increased stimulus spacing). Stimuli were presented with from 2.0 to 10.0 s between presentations. The genetically derived sequences for this experiment were generated in MATLAB through the Optimize Design toolbox by Wager and Nichols (2003).

The third stimulus presentation condition was an event-related design generated through a maximum length sequence (m-sequence) method. An m-sequence is a form of pseudo-random binary sequence that possesses extremely low autocorrelation. The low autocorrelation makes an m-sequence an optimal approach to maximize the ability to estimate the average hemodynamic response function for each condition. The caveat is that m-sequences will have lower statistical power than the block or genetic design conditions. Stimuli were presented with from 2.0 to 10.0 s between presentations. The m-sequences for this experiment were generated in MATLAB using an open-source script (Buracas & Boynton, 2002).

Image acquisition

Acquisition of BOLD fMRI data for all studies was completed on a Siemens Magnetom Trio 3.0-T whole body scanner with a 12-channel phased-array head coil for RF transmission and reception (Siemens Healthcare, Erlangen, Germany). The scanning parameters for the whole-brain T2* echo-planar imaging (EPI) sequence were as follows: 36 interleaved axial

slices (4 mm thick, 1-mm gap), TR = 2,000 ms, TE = 30 ms, flip angle = 90º, and 256 × 256 field of view. Four discarded acquisitions were used during the first 10 s of scanning to ensure magnetization equilibrium. Visual stimuli were projected onto a ground glass screen located at the head of the magnet bore by a digital projector. A mirror directly above the head coil allowed the subject to observe the projected image. Stimulus presentation was performed by the experiment-scripting program Psychophysics Toolbox (Brainard, 1997) and was synchronized to a transistor–transistor logic voltage trigger from the scanner.

Additional data set

To effectively investigate the multiple factors influencing fMRI reliability, we required additional data from an experiment with multiple testing sessions. Ideally, this data set would be acquired using the same scanner hardware, identical acquisition parameters, and a similar fMRI experimental design. We were fortunate to have one such data set readily available for our use. We used the data from this study ("Study 2") to examine the effects of increased test–retest interval on reliability measures.

The data for Study 2 were originally gathered by Miller, Donovan, Bennett, Aminoff, and Mayer (2012). In this study, they investigated the stability of intersubject variability in episodic recognition over the span of many months. The study included data from an episodic word recognition task that was a modified version of the event-related genetic design used in Study 1. It differed in having approximately 30 % more stimulus presentations to each subject. This was addressed by remodeling the data so as to only take into account the first 100 stimulus presentations during each session. The time between the test and retest samples used to investigate reliability was approximately six months. A random subset of 14 subjects was chosen from the full group of 16 subjects for this calculation, to match the sample size of the primary study.

Image-processing methods

Preprocessing for all studies was conducted using a similar processing pipeline in SPM5 (Wellcome Trust Centre for Neuroimaging, London, UK, www.fil.ion.ucl.ac.uk/spm), with all available software updates installed as of July, 2012. The images from the entire EPI time series were spatially realigned to the first image using a least squares approach with a six-parameter rigid-body affine transformation (Friston et al., 1995). The realigned images were then processed using the "unwarp" function to reduce the influence of residual movement-related variance on the BOLD signal intensity (Andersson, Hutton, Ashburner, Turner, & Friston, 2001). A coregistration step was then completed in which we used mutual information maximization with a six-parameter rigid-body affine transform to spatially align the EPI time-series data with a high-resolution T1 anatomical image. All images were then normalized into a standard 3-D stereotaxic space defined by the International Consortium for Brain Mapping (ICBM)-152 atlas space (Ashburner & Friston, 1999; Ashburner, Neelin, Collins, Evans, & Friston, 1997; Mazziotta, Toga, Evans, Fox, & Lancaster, 1995). This was done using parameters determined from a high-resolution T1 anatomical image. The normalization of the EPI images to the atlas template image was completed using a combination of 12-parameter linear affine transformation and 3 × 2 × 3 nonlinear 3-D discrete cosine transform. A 7th-degree B-spline was used as the interpolation method for creating normalized images. Finally, all images were smoothed with an 8-mm full-width-at-half-maximum (FWHM) isotropic Gaussian kernel.

We utilized the general linear model with restricted maximum likelihood estimation (ReML) for the statistical analysis of the preprocessed fMRI time-series data. Hemodynamic responses were modeled by a boxcar function convolved with the SPM5 canonical hemodynamic response for blocked stimulus presentation designs and by impulse functions convolved with the SPM5 canonical hemodynamic response for event-related stimulus presentation designs. A high-pass filter with a frequency cutoff of 128 s was used to remove low-frequency signal drift in the data. After estimation, a *t*-statistic contrast was used to evaluate task-versus-rest activity across the whole brain. In this contrast, all task trials were compared against an implicit rest condition generated by SPM. An additional contrast was generated to compare the target-versus-nontarget conditions in the episodic recognition task and the target-versus-nontarget conditions in the two-back task. The resulting volumes of *t*-statistic values for each subject were then entered into the reliability analyses and second-level group analyses.

Intraclass correlations

We chose to use the ICC, as defined by Shrout and Fleiss (1979), for the calculation of reliability. The ICC is similar to the Pearson correlation, but it is specialized for the correlation of data from the same class of information, such as *t*-statistic values. The following version of the formula was used to calculate ICC values:

$$\mathrm{ICC}(3,1) = (\mathrm{BMS} - \mathrm{EMS})/(\mathrm{BMS} + (k-1) \times \mathrm{EMS}).$$

One reason that we chose to calculate reliability using ICC values is that they can be interpreted very similarly to the Pearson correlation. A value of ICC=0 means that no relationship exists between the test and retest values, and a value of ICC=1 indicates perfect agreement between the test and retest values. The ICC measure is also a commonly used

metric of test–retest reliability across multiple disciplines and has been used frequently in the past to evaluate fMRI reliability (Aron, Gluck, & Poldrack, 2006; Bennett & Miller, 2010; Caceres et al., 2009; Eaton et al., 2008). This allows us to compare the results of this study against a broader range of the existing fMRI literature. Few other reliability measures have a comparable number of published results.

For the purposes of this study, we were interested in the stability of *t*-statistic values across subjects. This represents the likelihood that the statistic value observed for a region at Time 1 will be similar to the statistic value observed at Time 2. An in-house processing script run in the MATLAB computing environment (The MathWorks Inc., Natick, MA) was used to calculate ICC values. Functions from SPM5 were used to load the image volumes into memory for analysis.

Thresholding criteria

We generated two sets of voxels for use in subsequent processing. The first was an unthresholded set of voxels common to all subjects in the group. The only criterion for inclusion in this set was that a voxel be present in all subjects for all conditions of the experiment. The second voxel set was generated using only superthreshold voxels with positive changes in signal that were present in a second-level group analysis of each condition. The contrast estimates calculated from the test and retest sessions for each subject were entered into a second-level one-sample *t*-test to identify regions with significant positive activity during completion of the task. A corrected statistical threshold of probability of false discovery rate [p(FDR)]= .10 with a cluster extent threshold of k> 10 voxels was chosen to threshold the group result maps. This threshold was chosen to be a consistent but somewhat liberal threshold, to ensure an adequate sample of superthreshold voxels to use in the later analyses. Voxels that survived this analysis were selected for the set of thresholded data.

Histogram methods and distribution metrics

We plotted the ICC values of all voxels from the unthresholded set using a 100-bin histogram with a range of values between 1 and –1. The mean and median were calculated using functions from the MATLAB Statistics Toolbox (The MathWorks Inc., Natick, MA). To calculate the mode, we used the MATLAB Curve Fitting Toolbox (The MathWorks Inc., Natick, MA) to fit a mixture of two Gaussian functions to the histogram data points. The peak of this curve was taken to be the mode. This was done to address the irregular pattern of peak values across the bins, with several local maxima and minima. The curve-fitting strategy of multiple Gaussians was used to allow for the modeling of the main body of data while allowing for skew in the distribution of voxelwise ICC values.

Regionwise ICC values for statistical analysis

Whereas descriptive statistics were completed using data from all voxels, we used a regionwise approach for the statistical comparison of experimental factors. It would be improper to simply enter all available voxels into a subsequent statistical analysis, because each voxel is not an independent measure of local brain activity. Instead, each voxel has a high degree of spatial correlation with other nearby voxels. Ideally, we would be able to estimate the true number of independent signal sources across the brain, but that goal has proven to be quite difficult to achieve in practice (see Nichols & Hayasaka, 2003). To avoid estimating the true number of independent signal sources, we have instead decided to sample the mean ICC values across discrete regions of the brain, limiting the degrees of freedom to the number of separable anatomical regions.

For the analysis of unthresholded data, we chose to sample the mean ICC value across each of 90 cerebral regions defined by the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). This provides a mean regionwise ICC value for each anatomically defined region in the atlas. For the analysis of thresholded data, we also calculated the mean ICC values across regions, but the region was defined on the basis of the location of significant voxel clusters. After generating the group result maps for each condition, we then used an automated peak-search algorithm to identify separable clusters of voxels on the basis of statistical significance and total cluster size. Our criteria for this search were to identify clusters with p(FDR)< .10, voxel count k> 10 voxels, and a peak voxel more than 15 mm away from any other peak voxel. After these separable clusters were identified, a spherical region of interest (ROI) with a radius of 10 mm was sampled around the peak voxel of each cluster in order to obtain a mean ICC value for the local region. All mean regionwise ICC values were then used as the input for analyses of the thresholded data.

Results

Effects of experimental design, cognitive task, and contrast type

Using the 2 × 3 × 2 design of Study 1, we were able to investigate both main effects and interactions between cognitive task, experimental design, and contrast type. Because our analysis strategy provided mean ICC values from the same regions across all conditions, we utilized a repeated measures analysis of variance (ANOVA) to compare the influences of these factors on the mean regionwise ICC values. The regionwise mean ICC values were subjected to a Fisher

z-transformation prior to the analysis to ensure normality in the resulting data. The transformed, mean regionwise ICC values were then entered into the subsequent ANVOA analysis.

We found a significant effect of design on the ICC values, with the block design having higher overall values across subjects [Wilks' lambda= 0.132, $F(2, 88)= 288.52$, $p< .001$, $\eta_p^2= .868$]. A significant effect of task on ICC values was also apparent, with the two-back working memory task having higher overall ICC values across subjects [Wilks' lambda= 0.490, $F(1, 89)= 92.71$, $p< .001$, $\eta_p^2= .510$]. Finally, we observed a significant effect of contrast on ICC values, with the target-versus-nontarget contrast having higher overall ICC values across subjects [Wilks' lambda= 0.582, $F(1, 89)= 64.00$, $p< .001$, $\eta_p^2= .418$].

Some interactions between the principal factors were also significant. Task was found to interact with design [Wilks' lambda= 0.649, $F(2, 88)= 23.78$, $p< .001$, $\eta_p^2= .351$], and design was found to interact with contrast [Wilks' lambda= 0.394, $F(2, 88)= 67.82$, $p< .001$, $\eta_p^2= .606$]. Task did not interact with contrast, but the result did trend toward significance [Wilks' lambda= 0.956, $F(1, 89)= 4.09$, $p= .046$, $\eta_p^2= .044$].

See Table 1 for a complete listing of the voxelwise descriptive statistics, Fig. 1 for a series of histograms depicting the distributions of voxelwise ICC values for the task-versus-rest contrast, and Fig. 2 for histograms depicting the distribution of voxelwise ICC values for the target-versus-nontarget contrast.

Effects of experimental design, cognitive task, and contrast type after thresholding

We utilized a one-way ANOVA to compare the influences of cognitive task, experimental design, and contrast type on the mean regionwise ICC values obtained from superthreshold regions of interest. A repeated measures ANOVA could not be used, because the process of thresholding inevitably yields a different set of peak voxels for each task–design–contrast combination. The mean ICC values from each ROI were subjected to a Fisher *z*-transformation to ensure normality in the resulting data. The transformed, mean regionwise ICC values were then entered into the subsequent ANVOA analysis.

We noted a significant effect of design on ICC values after thresholding, with the block design having higher overall ICC values across subjects [$F(1, 164)= 43.07$, $p< .001$, $\eta_p^2= .082$]. A significant effect of task on ICC values also emerged, with the two-back working memory task having higher overall ICC values across subjects [$F(2, 164)= 7.30$, $p= .001$, $\eta_p^2= .208$]. Finally, the effect of contrast on ICC values was not significant [$F(1, 164)= 2.73$, $p= .101$, $\eta_p^2= .016$].

Significant interactions between the principal factors were also apparent: Task was found to interact with design [$F(2, 164)= 9.45$, $p< .001$, $\eta_p^2= .103$], and design to interact with contrast [$F(2, 164)= 5.11$, $p= .007$, $\eta_p^2= .059$]. Finally, task did not interact with contrast, but the result did trend toward significance [$F(1, 164)= 3.40$, $p= .067$, $\eta_p^2= .020$].

See Table 1 for a complete listing of the voxelwise descriptive statistics, and Fig. 3 for a series of histograms depicting the distribution of voxelwise ICC values for the thresholded task-versus-rest contrast. Histograms from the thresholded target-versus-nontarget contrast are not shown, as an insufficient quantity of active voxels were present for an effective visualization. The histograms for this analysis are nonetheless presented in the supplementary materials, as Fig. S1. We found no significant session effects in the group results at the $p(\text{FDR})< .05$ threshold.

Table 1 Descriptive statistics of ICC values observed in each condition

			Unthresholded				Thresholded			
Task	Design	Contrast	Mean	Median	Mode	*SD*	Mean	Median	Mode	*SD*
Episodic	Block	Task vs. Rest	.326	.358	.357	.305	.435	.487	.728	.351
	Genetic	Task vs. Rest	.299	.320	.407	.294	.350	.385	.524	.303
	m-Sequence	Task vs. Rest	.176	.167	.162	.334	.210	.247	.268	.369
	Block	Old vs. New	.364	.357	.381	.306	.381	.395	.414	.279
	Genetic	Old vs. New	.336	.370	.485	.316	.139	.095	.210	.360
	m-Sequence	Old vs. New	.138	.138	.141	.305	.154	.160	.178	.310
Two-Back	Block	Task vs. Rest	.465	.509	.599	.256	.461	.502	.595	.264
	Genetic	Task vs. Rest	.320	.340	.422	.273	.427	.491	.587	.323
	m-Sequence	Task vs. Rest	.283	.293	.343	.268	.409	.462	.529	.318
	Block	Old vs. New	.521	.577	.669	.259	.348	.353	.271	.301
	Genetic	Old vs. New	.351	.385	.487	.288	.239	.288	.517	.260
	m-Sequence	Old vs. New	.311	.335	.394	.296	.277	.303	.412	.281

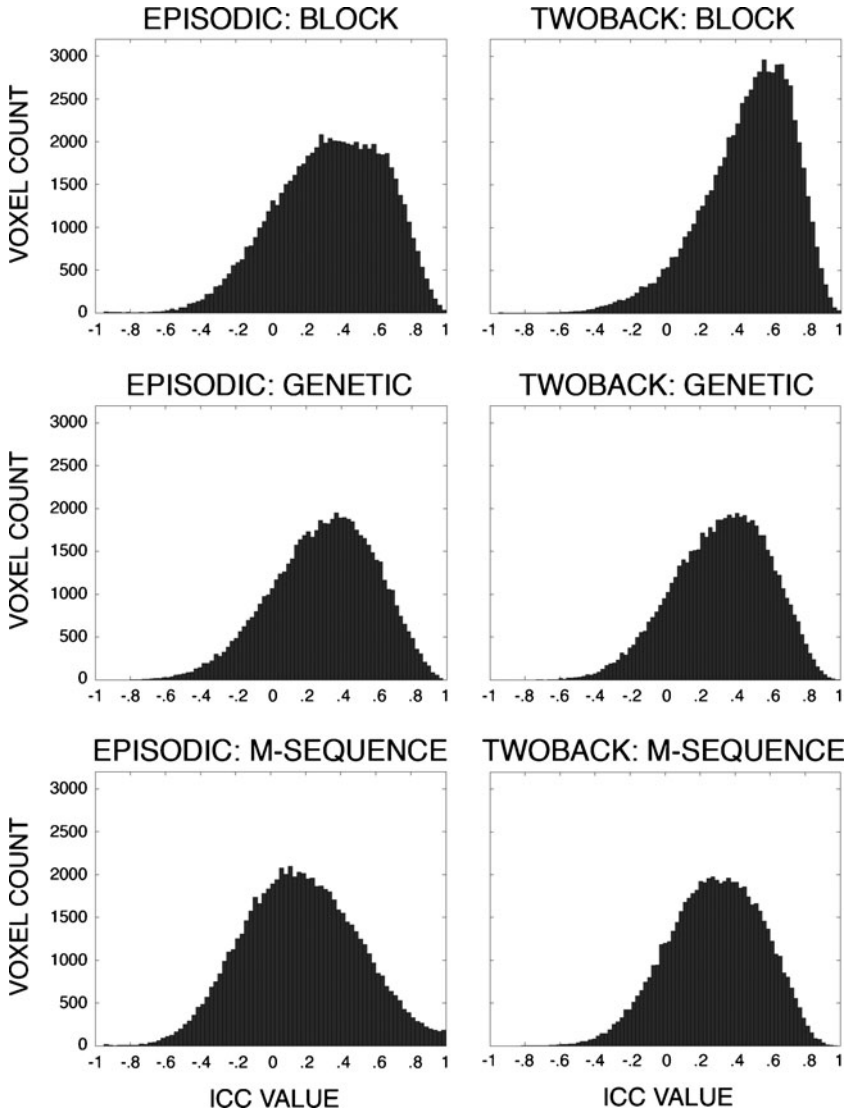

Fig. 1 Histograms depicting the distributions of ICC values in each condition for the unthresholded task-versus-rest contrast. All brain voxels common to all subjects are included in this analysis

Effects of test–retest interval on test–retest reliability

To evaluate the effects of short versus long test–retest intervals, we compared the data from the episodic recognition task in Study 1 and Study 2. This provided similar test–retest results, but over a six-month time interval in different groups of subjects. For the unthresholded data. we again used the approach of sampling the mean ICC value from each of 90 cerebral AAL atlas regions. For the thresholded results, we again used a peak-search algorithm to identify peak voxels that could be sampled with a spherical ROI. All ICC values were subjected to a Fisher *z*-transformation to ensure normality in the resulting data. A paired-samples *t*-test was used to test for mean differences in the ICC values for the unthresholded data. A two-tailed independent-samples *t*-test was used to formally test for mean differences in the ICC values for the thresholded data.

For the unthresholded task-versus-rest data, a significant difference was found between the two conditions, with the back-to-back scans having higher overall ICC values across subjects [$t(89) = 4.47$, $p < .001$, Cohen's $d = 0.498$]. For the thresholded task-versus-rest data, no significant difference was found between the two conditions, but the result did trend toward significance [$t(34) = 1.91$, $p = .065$, Cohen's $d = 0.655$]. Both effects were observed as a reduction in the mean, median, and mode for the distribution of ICC values at the six-month test–retest interval (see Fig. 4 and Table 2).

Relationship between significance and reliability

For each test–retest pairing, we calculated the Pearson product–moment correlation between significance, as measured by

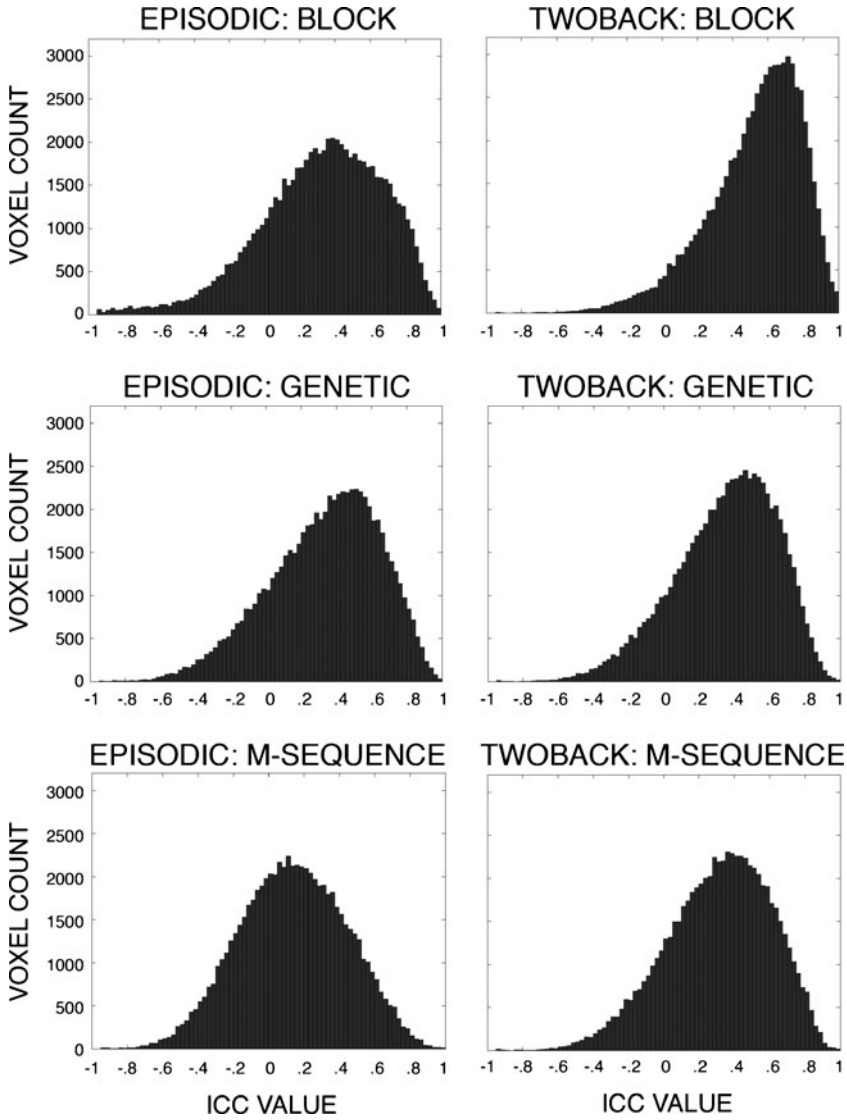

Fig. 2 Histograms depicting the distributions of ICC values in each condition for the unthresholded target-versus-nontarget contrast. All brain voxels common to all subjects are included in this analysis

the voxelwise *t*-statistic value, and reliability, as measured by the voxelwise Fisher *z*-transformed ICC value. The results for this correlation in the unthresholded task-versus-rest contrast ranged from $r = .02$ to .33, with a mean of .17. This relationship remained similar even when only statistically significant voxels were used in the correlation. The values for the correlation calculation in superthreshold voxels of the same contrast ranged from –.06 to .21, with a mean of .07. Figure 5 shows the group results and ICC results for the task-versus-rest contrast of the episodic recognition task. Figure 6 shows the group results and ICC results for the task-versus-rest contrast of the two-back working memory task. The group results and ICC results for the target-versus-nontarget contrast are included in the supplementary materials. For a full breakdown of the correlation values, please see Table S1 in the supplemental materials.

Discussion

Reliability is a critical metric by which to judge our scientific tools and methods. A major contribution of the present study relative to the established literature has been the examination of reliability across multiple experimental factors simultaneously. In this study, we have demonstrated that experimental design, cognitive task, contrast type, thresholding, and test–retest interval can all influence the reliability of *t*-statistic values across the whole brain. Furthermore, these factors can interact to produce increased or reduced reliability of the results, depending on the choices of task and design.

Effects of cognitive task

We found that cognitive tasks can vary with regard to their reliability. Specifically, in our results we found that the two-

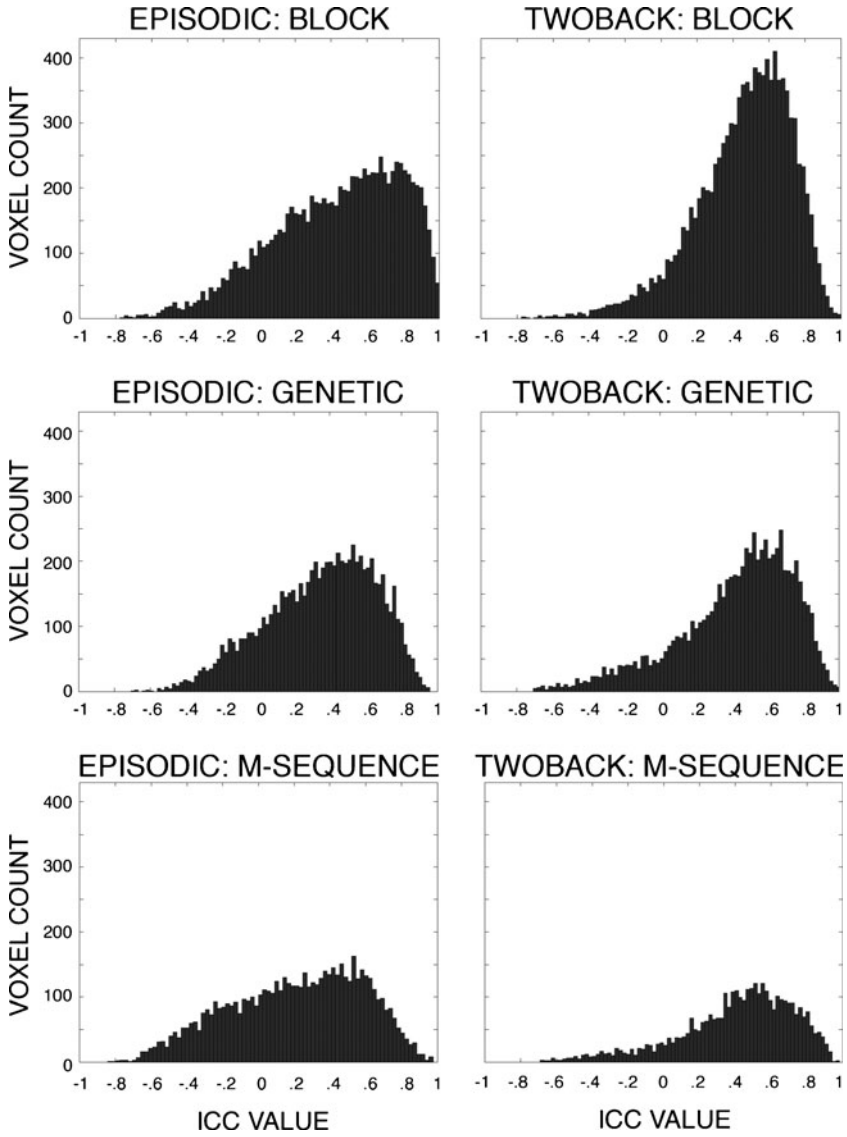

Fig. 3 Histograms depicting the distributions of ICC values in each condition for the thresholded task-versus-rest contrast. Only the set of superthreshold voxels present in each condition are included in the analysis

back working memory task had a higher overall reliability than did the episodic recognition memory task. This was true within the restricted set of thresholded results and within the unthresholded results including all brain voxels. No two tasks require the same pattern of regional brain activity, and it is logical that these varying task demands could influence the reliability of the task results (Miller et al., 2012).

Effects of experimental design

We found that the experimental design of a study can also impact the reliability of the results. For our data, block designs tended to have higher reliability than did either the event-related genetic design or the event-related m-sequence design. Although we observed some relationship between the *t*-statistic values and the ICC reliability values, the correlation could be said to be weak at best.

Effects of contrast type

Although it is not typical to use a task-versus-rest contrast in an empirical fMRI study, it is still instructive to examine how its reliability may be different from a condition-versus-condition contrast. As such, we found that the type of contrast chosen by the investigator can influence the reliability of the results. The target-versus-nontarget contrast tended to have higher reliability values than did the task-versus-rest contrast in our data set.

Effects of test–retest interval on reliability

We found that reliability values tend to decay with increasing test–retest interval. Back-to-back scanning runs for an unthresholded, event-related episodic word recognition task with a genetic design had a mean ICC value of .299 across

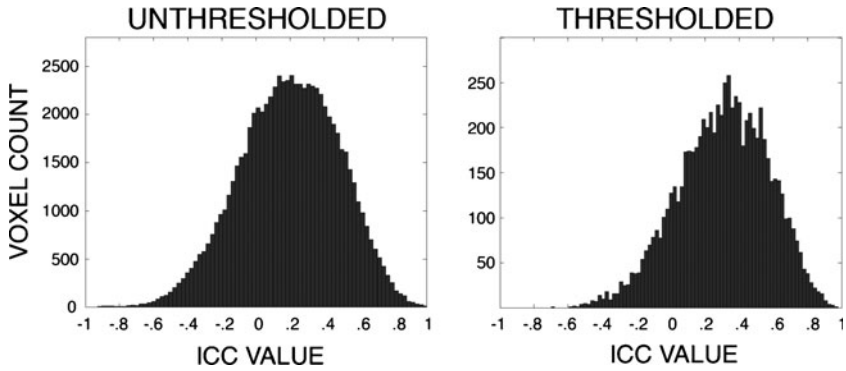

Fig. 4 Histograms depicting the distributions of ICC values at a six-month test–retest interval. Both unthresholded and thresholded values are depicted

voxels, whereas equivalent scanning runs that were six months apart had a mean ICC value of .197 across voxels. It is parsimonious that longer test-retest intervals would have reduced ICC values. It is likely that an increased number of factors that contribute to the error variance would be present as the test-retest interval increases. However, there are certainly exceptions to this rule; for instance, Aron, Gluck, and Poldrack (2006) found exceptionally high regional ICC values after an extended 12-month test–retest interval.

The interaction of multiple factors

We found that the principal factors shown to influence the reliability of a result can also interact to create higher or lower reliability. In our study, the factors of Experimental Design and Cognitive Task interacted with a medium effect size, leading to higher ICC values in the two-back block design condition. This lends greater importance to understanding how these separate factors combine to influence fMRI reliability: If factors can interact synergistically, as may have been the case here, it is critical to understand the mechanisms behind how these factors can potentially reinforce each other.

Relationship between significance and reliability

Across all test–retest examinations, the relationship between voxel significance and voxel reliability was somewhat small. This speaks against the intuitive view that the test–retest reliability of significant voxels will necessarily be higher than that of subthreshold voxels. For some sets of results, the distribution of ICC values across the pool of significant voxels was virtually identical to the distribution of ICC values across the whole brain. Although significant voxels had a greater probability of high ICC values, we did not find a strong relationship between significance and reliability. This finding is similar to those of both Caceres et al. (2009) and Fliessbach et al. (2010).

Overall, all of our principal factors of investigation were significant, with some factors interacting with each other. It would be easy to simply conclude that all factors affect reliability, as this would be a true statement, given the results. However, although we took a number of steps to ensure that most of the many influential factors would be equivalent across conditions, we cannot exclude the possibility that other factors (or interacting factors) might have influenced the observed results. As we stated in the introduction, each experimental combination of task, design, scanner, image-processing pipeline, and analysis strategy has a reliability value that is potentially unique unto itself. This could mean that another study, with parameters identical to our own, could have a very different pattern of results as other factors interacted with those manipulated in this study. In this context, it is perhaps more important to examine the relative effect sizes of each principal factor. This yields a framework to further compare and contrast the relative contributions of each factor to the reliability of the results, as measured by ICC.

In our results, we reported effect sizes primarily in terms of η_p^2. Partial eta-squared represents the amount of variability in the dependent variable that can be explained after the other

Table 2 Descriptive statistics of ICC values obtained at a test–retest intervals of 20 min and 6 months

Task	Design	Contrast		Unthresholded				Thresholded			
				Mean	Median	Mode	*SD*	Mean	Median	Mode	*SD*
Episodic	Genetic	Task vs. Rest	20 min	.299	.320	.407	.294	.350	.385	.524	.303
			6 months	.197	.206	.212	.275	.305	.321	.327	.264

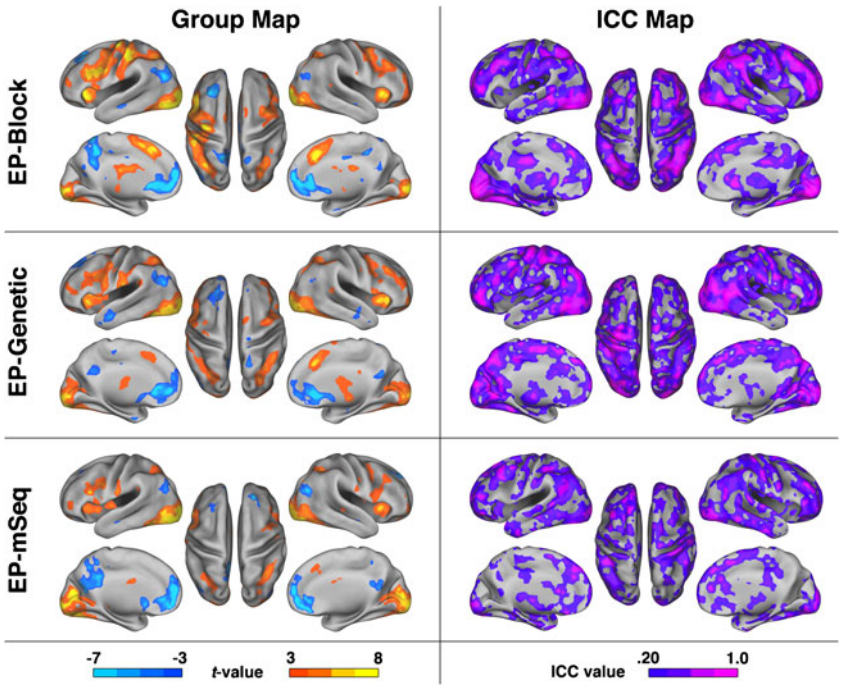

Fig. 5 Surface renderings of the thresholded second-level group results and the test–retest ICC values for the task-versus-rest contrast of the episodic recognition task. The group maps were each thresholded at a level of $t > 3.0$ for consistent visualization across conditions. The ICC maps were arbitrarily thresholded at a level of ICC> .20 to provide for better visualization of the areas with low and high ICC values

principal factors have been accounted for. It is calculated by the following equation:

$$\eta_p^2 = SS_{effect}/(SS_{effect} + SS_{error}).$$

This is a useful metric for effect size, because it is somewhat analogous to r^2 and can be thought of as a percentage of the variance accounted for. Interpretations of η_p^2 have been proposed such that .01 indicates a small, .06 a medium, and .14 a large effect (Cohen, 1988).

For the unthresholded data, the largest effect observed was for the Design factor ($\eta_p^2 = .868$). The factors Contrast Type and Task had smaller effect sizes that were also in the range of large effects, with contrast type having a η_p^2 of .418 and task a η_p^2 of .510. When examining the effect sizes of the observed

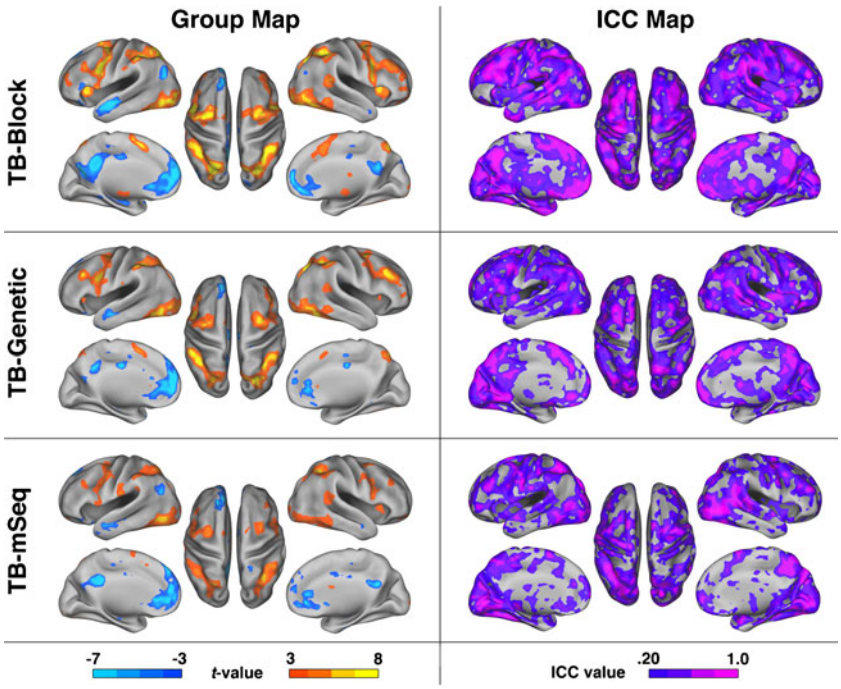

Fig. 6 Surface renderings of the thresholded second-level group results and the test–retest ICC values for the task-versus-rest contrast of the two-back task. The group maps were each thresholded at a level of $t > 3.0$ for consistent visualization across conditions. The ICC maps were arbitrarily thresholded at a level of ICC> .20 to provide for better visualization of the areas with low and high ICC values

interactions, the Design × Task interaction had a large effect size of .351, and the Design × Contrast interaction had a large effect size of .394.

Although the principal factors for the thresholded data were also significant, the effect sizes were far lower, relative to the unthresholded data: None of the principal factors or interactions could explain much more than 1 % of the variability in the sample. This was likely due to the different statistical tests involved, with the unthresholded test using a repeated measures ANOVA over the same pool of voxels, and the thresholded data using a one-way ANOVA over varying pools of significant voxels. However, the pattern of results diverged from that of the unthresholded results, with the principal factor of Task having the largest effect size, followed by Design. Conversely, the pattern of results for the interactions mirrored that of the unthresholded values, with Design × Task having a larger effect size than Design × Contrast.

Our findings fit in well with the established literature on test–retest reliability in fMRI. Bennett and Miller (2010) conducted a large-scale meta-analysis of the existing test–retest reliability literature. They found that the results of test–retest fMRI reliability studies using intraclass correlation as a metric had a mean ICC value of .50. This was calculated across 15 previous studies that had used ICC to evaluate test–retest fMRI reliability. The mean values from the comparisons examined in the present study are at or below this value. This is not to say that the results of every past, present, or future fMRI study will be reliable at this mean level. Some studies have shown results higher than this value, such as those of Aron et al. (2006), and some necessarily have been lower, such as those of Kong et al. (2007). Still, our finding does give an indication of the test–retest reliability of fMRI in general terms and is the start toward building an understanding of fMRI reliability values across cognitive tasks and experimental designs. It should also be noted, however, that this is only one small facet of characterizing fMRI reliability. For example, Raemaekers et al. (2007) found that, whereas group fMRI results can be highly reliable, variability in the reliability of statistical maps from individual subjects was high. Those researchers also explored the regional variability in reliability across the brain, whereas this study focused on the reliability of the brain as a whole. Future studies that examine reliability in terms of local ROIs may obtain values dramatically different from our whole-brain results, simply due to the scope of measurement.

Intraclass correlation is a frequently used measure of test–retest reliability. However, this does not mean that it is without weaknesses. One major shortcoming of the ICC approach is the simultaneous inclusion of between- and within-subjects variability. For the exact same degree of within-subjects variability, represented by the stability of test–retest values, different ICC values may result, depending on the degree of between-subjects variability. For example, many clinical populations possess an increased level of between-subjects variability that can influence ICC values. One documented example was a sample of stroke victims performing a visuomotor task. Kimberly, Khandekar, and Borich (2008) found that the estimated ICC values for clinical patients were higher than those for normal controls on the same task. It should also be noted that, whereas the ICC does include intersubject variability in the calculation, the results cannot necessarily be construed as a measure of repeatability with a new set of subjects (Caceres et al., 2009). This is important, as the preservation of individual differences in brain activity over time can be a critical variable in many fMRI investigations (Miller et al., 2002, 2009).

Summary

Quantifying the reliability of fMRI data is notoriously difficult. As we have demonstrated, not only do most factors affect the reliability of fMRI data, but they do so to varying degrees of influence. Even the amplitude of the BOLD response itself can vary over sessions (Raemaekers et al., 2012). Our results speak strongly in favor of future studies reporting reliability or internal consistency values whenever possible. Potentially, this could be a simple matter for most studies that acquire several runs of data during scanning: Simple split-half evaluations of the data using intraclass correlation could give future readers a more thorough understanding of how robust the reported results are. Other measures of internal consistency throughout a scan could also be potentially useful to a reader of a study's Results section. One recommendation that we support is for major neuroimaging software packages to include these types of calculations automatically as part of their analysis (Braver et al., 2010).

The body of emerging evidence indicates that fMRI reliability is a highly variable construct, influenced by factors both within and outside the control of the investigator. Given this context, it becomes imperative to determine what the expected reliability of fMRI results may be and what factors can influence that reliability. This is especially true when increased understanding of these factors can lead to more highly optimized experiments with more reliable results.

Author note This work on fMRI reliability was supported by the Institute for Collaborative Biotechnologies, through Contract No. W911NF-09-D-0001 from the U.S. Army Research Office.

References

Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *NeuroImage, 13,* 903–919.

Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test–retest reliability of functional MRI in a classification learning task. *NeuroImage, 29,* 1000–1006.

Ashburner, J., & Friston, K. J. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping, 7,* 254–266.

Ashburner, J., Neelin, P., Collins, D. L., Evans, A., & Friston, K. (1997). Incorporating prior knowledge into image registration. *NeuroImage, 6,* 344–352.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191,* 133–155.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Braver, T. S., Cole, M. W., & Yarkoni, T. (2010). Vive les differences! Individual variation in neural mechanisms of executive control. *Current Opinion in Neurobiology, 20,* 242–250. doi:10.1016/j.conb.2010.03.002

Brodersen, K. H., Wiech, K., Lomakina, E. I., Lin, C. S., Buhmann, J. M., Bingel, U., & Tracey, I. (2012). Decoding the perception of pain from fMRI using multivariate pattern analysis. *NeuroImage, 63,* 1162–1170.

Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., & Potkin, S. G. (2011). Multisite reliability of cognitive BOLD data. *NeuroImage, 54,* 2163–2175.

Buracas, G. T., & Boynton, G. M. (2002). Efficient design of event-related fMRI experiments using M-sequences. *NeuroImage, 16,* 801–813.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Eaton, K. P., Szaflarski, J. P., Altaye, M., Ball, A. L., Kissela, B. M., Banks, C., & Holland, S. K. (2008). Reliability of fMRI for studies of language in post-stroke aphasia subjects. *NeuroImage, 41,* 311–322.

Fliessbach, K., Rohe, T., Linder, N. S., Trautner, P., Elger, C. E., & Weber, B. (2010). Retest reliability of reward-related BOLD signals. *NeuroImage, 50,* 1168–1176.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. (1995). Spatial registration and normalization of images. *Human Brain Mapping, 2,* 165–189.

Gradin, V., Gountouna, V. E., Waiter, G., Ahearn, T. S., Brennan, D., Condon, B., & Steele, J. D. (2010). Between- and within-scanner variability in the CaliBrain study *n*-back cognitive task. *Psychiatry Research, 184,* 86–95.

Harrington, G. S., Buonocore, M. H., & Farias, S. T. (2006a). Intrasubject reproducibility of functional MR imaging activation in language tasks. *American Journal of Neuroradiology, 27,* 938–944.

Harrington, G. S., Farias, S. T., Buonocore, M. H., & Yonelinas, A. P. (2006b). The intersubject and intrasubject reproducibility of FMRI activation during three encoding tasks: Implications for clinical applications. *Neuroradiology, 48,* 495–505.

Havel, P., Braun, B., Rau, S., Tonn, J. C., Fesl, G., Bruckmann, H., & Ilmberger, J. (2006). Reproducibility of activation in four motor paradigms: An fMRI study. *Journal of Neurology, 253,* 471–476.

Kimberley, T. J., Khandekar, G., & Borich, M. (2008). fMRI reliability in subjects with stroke. *Experimental Brain Research, 186,* 183–190.

Kong, J., Gollub, R. L., Webb, J. M., Kong, J. T., Vangel, M. G., & Kwong, K. (2007). Test–retest study of fMRI signal change evoked by electroacupuncture stimulation. *NeuroImage, 34,* 1171–1181.

Koolschijn, P. C., Schel, M. A., de Rooij, M., Rombouts, S. A., & Crone, E. A. (2011). A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test–retest reliability from childhood to early adulthood. *Journal of Neuroscience, 31,* 4204–4212.

Liu, T. T. (2004). Efficiency, power, and entropy in event-related fMRI with multiple trial types: Part II. *Design of experiments. NeuroImage, 21,* 401–413.

Liu, T. T., & Frank, L. R. (2004). Efficiency, power, and entropy in event-related fMRI with multiple trial types: Part I. *Theory. NeuroImage, 21,* 387–400.

Matthews, P. M., Honey, G. D., & Bullmore, E. T. (2006). Applications of fMRI in translational medicine and clinical practice. *Nature Reviews Neuroscience, 7,* 732–744.

Maus, B., van Breukelen, G. J., Goebel, R., & Berger, M. P. (2010). Robustness of optimal design of fMRI experiments with application of a genetic algorithm. *NeuroImage, 49,* 2433–2443.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). *NeuroImage, 2,* 89–101.

Miller, M. B., Donovan, C. L., Bennett, C. M., Aminoff, E. M., & Mayer, R. E. (2012). Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *NeuroImage, 59,* 83–93.

Miller, M. B., Donovan, C. L., Van Horn, J. D., German, E., Sokol-Hessner, P., & Wolford, G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *NeuroImage, 48,* 625–635.

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., & Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience, 14,* 1200–1214.

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research, 12,* 419–446.

Raemaekers, M., du Plessis, S., Ramsey, N. F., Weusten, J. M. H., & Vink, M. (2012). Test–retest variability underlying fMRI measurements. *NeuroImage, 60,* 717–727.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R. J. A., Kahn, R. S., & Ramsey, N. F. (2007). Test–retest reliability of fMRI activation during prosaccades and antisaccades. *NeuroImage, 36,* 532–542.

Rau, S., Fesl, G., Bruhns, P., Havel, P., Braun, B., Tonn, J. C., & Ilmberger, J. (2007). Reproducibility of activations in Broca area with two language tasks: A functional MR imaging study. *American Journal of Neuroradiology, 28,* 1346–1353.

Sato, J. R., Hoexter, M. Q., Fujita, A., & Rohde, L. A. (2012). Evaluation of pattern recognition and feature extraction methods in ADHD prediction. *Frontiers in Systems Neuroscience, 6,* 68.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Soltysik, D. A., Thomasson, D., Rajan, S., Gonzalez-Castillo, J., DiCamillo, P., & Biassou, N. (2011). Head-repositioning does not reduce the reproducibility of fMRI activation in a block-design motor task. *NeuroImage, 56,* 1329–1337.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage, 15,* 273–289. doi:10.1006/nimg.2001.0978

Wager, T. D., & Nichols, T. (2003). Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *NeuroImage, 18,* 293–309.

Waldvogel, D., van Gelderen, P., Immisch, I., Pfeiffer, C., & Hallett, M. (2000). The variability of serial fMRI data: Correlation between a visual and a motor task. *NeuroReport, 11,* 3843–3847.

Yetkin, F. Z., McAuliffe, T. L., Cox, R., & Haughton, V. M. (1996). Test–retest precision of functional MR in sensory and motor task activation. *American Journal of Neuroradiology, 17,* 95–98.

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., & Strother, S. C. (2009). Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magnetic Resonance Imaging, 27,* 264–278.