How reliable are the results from fMRI?

Craig M. Bennett[1], Christa-Lynn Donovan[2], Scott Guerin[3], Michael Miller[1]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Stanford University, Stanford, CA; [3] Harvard University, Cambridge, MA

INTRODUCTION

Functional neuroimaging using fMRI has proven to be a popular and effective method of investigating regional brain activity in vivo. A large body of papers have been published that use fMRI, but only a small number of papers have investigated the reliability of fMRI results. By understanding the many factors that influence reliability we can potentially increase the measurement precision of both group results and inter-individual differences. With this work we take important steps forward in understanding the test-retest reliability of fMRI results.

Our overall hypothesis was that factors that increase statistical power or reduce intra-individual variability would increase the reliability of results. To test this hypothesis we examined reliability according to cognitive task, experimental design, and test-retest interval. We also examined the broader landscape of fMRI reliability by reporting findings from a review of the existing test-retest reliability literature.

METHODS

The main fMRI experiment was a 2x2 design with principal factors of task type (episodic word recognition and two-back working memory) and stimulus presentation strategy (block and event-related designs). Fourteen participants completed eight functional imaging runs, two runs per condition, to yield 100 stimulus presentations per condition. The test-retest interval for each condition was approximately 20 minutes.

A general linear model was estimated and task-versus-rest contrasts were constructed for each run independently. Reliability of these contrasts was evaluated using a voxelwise intra-class correlation between test and retest runs (ICC; Shrout and Fleiss, 1979). The ICC values for voxels across the whole brain and for significant voxels were then calculated. Pilot testing was completed before scanning took place to ensure equivalent behavioral performance in each condition.

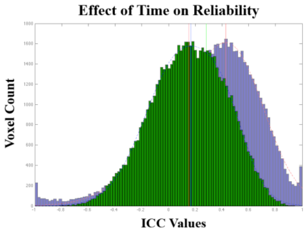
We compared the reliability of the above experiment to the reliability of another study (Donovan and Miller, 2009) completed using the same Siemens 3T MR scanner. The study was included to examine the reliability of a similar event-related episodic word recognition task across a six-month test-retest interval. The calculation of voxelwise ICC values was completed in the same manner as the above experiment.

To find papers for the literature review we searched for “test-retest fMRI” using the NCBI PubMed database (www.pubmed.gov). This search yielded a total of 183 papers, 37 of which used fMRI as a method of investigation, used a general linear model to compute their results, and provided test-retest measures of reliability. To broaden the scope of the search we then examined the reference section of these papers to look for additional publications. The total number of papers included was 63. We only report the reliability values from studies that used either intra-class correlation or voxel overlap statistics.

META-ANALYSIS RESULTS

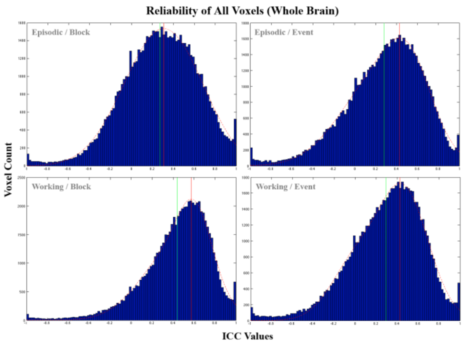
The mean ICC value across 13 published studies was found to be 0.50. The mean overlap of significant voxels from test and retest sessions was found to be 0.48 as calculated using the Dice coefficient equation, or 0.33 as measured by the Jaccard overlap equation. This means that, out of all significant voxels in a test-retest session, on average only 33% would be the same in each image.

RELIABILITY OVER TIME



We also found that an extended test-retest interval reduced the reliability of fMRI results for event-related episodic long-term memory task. Figure 3 (below) is a histogram illustrating voxelwise ICC values for 20 minute (blue) and six-month (red) test-retest intervals. The mean ICC value of episodic runs separated by 20 minutes was 0.28 while a six-month test-retest period had a mean ICC value of 0.15.

RELIABILITY BY TASK/DESIGN



We found that multiple factors can have an interacting influence on fMRI reliability. Figure 1 (above) shows histograms depicting voxelwise ICC values from each condition. The mean ICC value across the whole brain for each condition was: 0.28 for episodic/event, 0.27 for episodic/block, 0.29 for working/event, and 0.45 for working/block. This indicates that different task/design pairings can have varying levels of reliability.

IMPACT OF THRESHOLDING

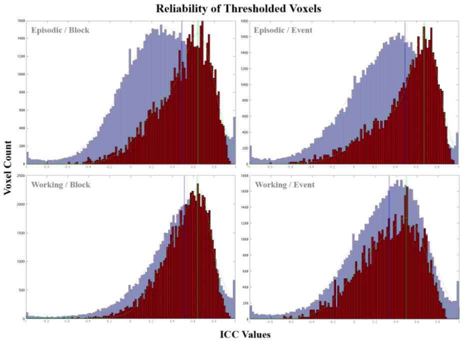


Figure 2 (above) shows a similar set of histograms, but calculated using only significant voxels from each analysis ($t > 3.2$, $p < 0.001$, uncorrected). The mean ICC value of significant voxels in the working memory conditions was approximately the same (working/event: 0.33, working/block: 0.51), while significant voxels in the episodic memory conditions had increased reliability (episodic/event: 0.49, episodic/block: 0.48). This indicates that the reliability of significant voxels may not be consistent from study-to-study.

CONCLUSIONS

Reliable measurement is the foundation on which scientific investigation is based. Few studies have attempted to directly evaluate how reliability varies according to experimental design or task. The results of our investigation show that numerous factors can impact the reliability of fMRI. It also illustrates that these factors can interact to raise or lower the reliability of an experiment. Our investigation demonstrates that the average reliability of fMRI investigations may be lower than many investigators implicitly assume (Vul et al., 2009; Lieberman et al., 2009).

REFERENCES

Donovan, C.L. and Miller, M.B., 2009. Longitudinal assessment of individual differences in fMRI. 16th Annual Meeting of the Cognitive Neuroscience Society. San Francisco, CA.

Lieberman, M.D., Berkman, E.T., Wager, T.D., 2009. Correlations in social neuroscience aren't voodoo: Commentary on Vul et al. (2009). Perspectives on Psychological Science 4.

Shrout, P., Fleiss, J., 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86, 420-428.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on Psychological Science 4.

Supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-D-0001 from the U.S. Army Research Office.