NeuroImage 50 (2010) 572–576

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Comments and Controversies

Avoiding non-independence in fMRI data analysis: Leave one subject out

Michael Esterman *, Benjamin J. Tamber-Rosenau, Yu-Chin Chiu, Steven Yantis

Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

ARTICLE INFO

Article history:
Received 27 May 2009
Revised 9 October 2009
Accepted 18 October 2009
Available online 16 December 2009

ABSTRACT

Concerns regarding certain fMRI data analysis practices have recently evoked lively debate. The principal concern regards the issue of non-independence, in which an initial statistical test is followed by further non-independent statistical tests. In this report, we propose a simple, practical solution to reduce bias in secondary tests due to non-independence using a leave-one-subject-out (LOSO) approach. We provide examples of this method, show how it reduces effect size inflation, and suggest that it can serve as a functional localizer when within-subject methods are impractical.

Two recent commentaries have suggested that certain widely used functional MRI data analysis practices may have led to exaggerated or even baseless claims (Vul et al., 2009; Vul and Kanwisher, in press). The principal concern regards the issue of statistical non-independence, in which data selected on the basis of an initial statistical test are subjected to one or more further (non-independent) statistical test(s). One claimed consequence is that reported effect sizes[1] can be misleadingly inflated, perhaps to "impossibly high" levels. While Vul et al. (2009) dealt specifically with individual-differences correlations, Vul and Kanwisher (in press) and Nichols and Poline (2009) make clear that these issues apply equally to other measures of effect size, which often depend upon beta weights from a general linear model (GLM) used in many fMRI analyses. Responses to Vul et al. (2009) range from the assertion that in most cases, a second non-independent statistical test was not actually performed but that an effect size was reported from brain regions identified by the initial statistical analysis (e.g. Lieberman et al., 2009), to the assertion that the non-independence error leads to so much confusion that a multistep framework should be reconsidered entirely and perhaps discarded (Lindquist and Gelman, 2009). Another recent article (Kriegeskorte et al., 2009) addressed the problem of non-independence in a broader range of neuroimaging analyses and suggested a step-by-step procedure for assessing and avoiding these kinds of errors.

In this report, we propose a simple and practical solution to the problem of non-independence that addresses the potential inflation of effect size and reduces the degree to which any secondary statistics are dependent on the initial statistical test. Our approach pertains to situations in which (1) group brain imaging data reveal one or more regions of interest associated with an experimental manipulation or a correlation with behavior, questionnaire data, or genotype, and (2) subsequent investigation of the discovered region(s) is carried out (e.g., effect size evaluation or multivariate pattern analysis). These are standard fMRI data analysis techniques and are performed in most or all of the work critiqued by Vul and colleagues. Our method is especially useful when a within-subject independent functional localizer is desired, but impractical.

The method employs a leave-one-subject-out (LOSO) cross-validation procedure in which a single subject is iteratively left out of the first-stage group analysis (here, a GLM with subject as random factor). The group GLM defines region(s) of interest which are applied to the data collected from the subject left out. Subsequent analysis is then carried out using the left-out subject's data (e.g., beta weights, raw signals, etc.) that are extracted from these region(s), and the procedure is then repeated for each subject. The GLM from the remaining subjects thus serves as an independent localizer for the subject left out (e.g. Esterman et al., 2009). The idea of cross-validation is not novel; in fact, several of the commentaries cited earlier suggest similar ideas (e.g., Kriegeskorte et al., 2009). However, others propose a more labor intensive within-subject leave-one-run-out cross-validation (where independence is arguably less assured, since data are selected using a region defined with data taken from the very same subject), or a potentially less-sensitive, between-subjects split-half cross-validation in which half of the data from each subject are used for region definition, and secondary analysis is carried out using data extracted from the other half. In the following sections, we illustrate the LOSO technique and show how it greatly reduces the effects of the non-independence error.

Methods

The LOSO technique is demonstrated with two data sets, one a block design, and the other a slow event-related design. Both data sets came from a study of top–down effects in category-specific visual processing; the results of that study and further methodological details are reported elsewhere (Esterman and Yantis, 2009).

* Corresponding author.
E-mail addresses: esterman@jhu.edu, esterman@gmail.com(M. Esterman).

[1] The effect sizes discussed here and in the work of Vul and colleagues (2009, in press) are *not* standardized measures of effect size (e.g., η^2). Instead, they are the absolute magnitude of the effect (e.g., correlation or beta weight) and thus are not comparable between studies.

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.10.092

Participants

A group of nine graduate and undergraduate students participated in a single fMRI session in which both data sets were collected. All participants provided informed consent as approved by the Johns Hopkins Medicine Institutional Review Board.

fMRI acquisition

MRI scanning was carried out with a Philips Intera 3T scanner in the F. M. Kirby Research Center for Functional Brain Imaging at the Kennedy Krieger Institute, Baltimore, MD. Anatomical images were acquired using an MP-RAGE T1-weighted sequence that yielded images with 1 mm isotropic voxels (TR = 8.1 ms, TE = 3.7 ms, flip angle = 8°, time between inversions = 3 s, inversion time = 738 ms). Whole-brain echo-planar functional images (EPI) were acquired with an 8-channel SENSE (MRI Devices, Inc., Waukesha, Wisconsin) parallel-imaging head coil in 40 transverse slices (TR = 2000 ms, TE = 35 ms, flip angle = 90°, matrix = 64 × 64, FOV = 192 mm, slice thickness = 3 mm, no gap). Neuroimaging data were analyzed using BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Functional data were slice acquisition time and motion corrected and then temporally high-pass filtered to remove frequency components in the time course of a run for each voxel, with a frequency of 3 cycles per run or fewer. This removed gradual drift in signal not associated with the critical task events (which occurred at roughly 12–24 cycles per run). Each subject's EPI volumes were all coregistered to that subject's anatomical scan. Spatial smoothing was applied (4-mm FWHM Gaussian kernel). Anatomical and functional images were Talairach-transformed and resampled into 3-mm isotropic voxels.

Data Set 1: blocked face or house 1-back matching task

Paradigm

In the block-design task, 12 faces and 12 houses were presented, either unscrambled or with randomly scrambled phases, for 15-s blocks at a rate of 1 image per second. Participants performed a 1-back matching task with four conditions: faces, houses, scrambled faces, and scrambled houses; each block was presented 4 times per run. Participants completed 3 or 4 runs of the block-design task (one participant only performed a single run).

General linear models

Face-selective activation maps were defined by the contrast between the face block and all other blocks (houses, scrambled faces, and scrambled houses). Face, house, scrambled face, and scrambled house block regressors were defined as a canonical hemodynamic response function (Boynton et al., 1996) convolved with a boxcar function of the block duration. A group GLM was performed with subjects as a random effect. In this full group GLM ($n = 9$), a minimum cluster size of 8 contiguous voxels was adopted to correct for multiple comparisons, yielding a whole-brain corrected statistical threshold of $\alpha < 0.01$ ($t = 3.355$, nominal $p < 0.01$) determined by a cluster threshold estimator plug-in implemented in BrainVoyager. Effect sizes for a given subject derived from this group GLM were non-independent, as all subjects were included in the GLM, and thus potentially inflated. In contrast, the independent analysis was accomplished by performing nine separate LOSO GLMs, one with each subject left out. For each of the nine LOSO GLMs ($n = 8$ subjects per GLM), two thresholds were used. First, we preserved the original nominal p value (0.01), thereby increasing the critical t value (3.499) because of the reduction in the degrees of freedom. Second, we used a more liberal threshold – the original t value – reducing the nominal p value ($p = 0.01267$). We call these the "conservative independent" and "liberal independent" tests, respectively. We chose to maintain the cluster threshold for the LOSO GLMs.

Regions of interest

Four regions were defined in the full group GLM to test the LOSO method (see Fig. 1A). These were two face-selective regions in the middle fusiform gyrus, consistent with the left and right fusiform face area (FFA; Sergent et al., 1992; Allison et al., 1994; Puce et al., 1995; Kanwisher et al., 1997) and another face-selective region in the right superior temporal sulcus (STS; Puce et al., 1998; Haxby et al., 1999). Finally, there was a fourth face-selective region in the right inferior occipital cortex (OFA; Gauthier et al., 2000; Rossion et al., 2003).

In each LOSO GLM, for each significant cluster of voxels, we identified the region or anatomical neighborhood of the cluster using a standard Talairach atlas as well as previous work describing face-selective cortex. This process is akin to defining ROIs using within-subject independent localizers. Importantly, these neighborhoods were not constrained by the voxels identified in the full group analysis. We then examined all LOSO ROIs to determine whether activity in a given anatomical neighborhood (i.e., right fusiform gyrus or right superior temporal sulcus) was consistent across all LOSO GLMs. This method could theoretically lead to LOSO ROIs that do not overlap, though in practice, it tends to lead to LOSO ROIs that overlap partially, as illustrated in Fig. 2 for the rFFA. Further, any fold of LOSO could discover multiple clusters within a broader region. If, for instance, two clusters were identified, the experimenter could consider signal from voxels in both subregions together, or if there

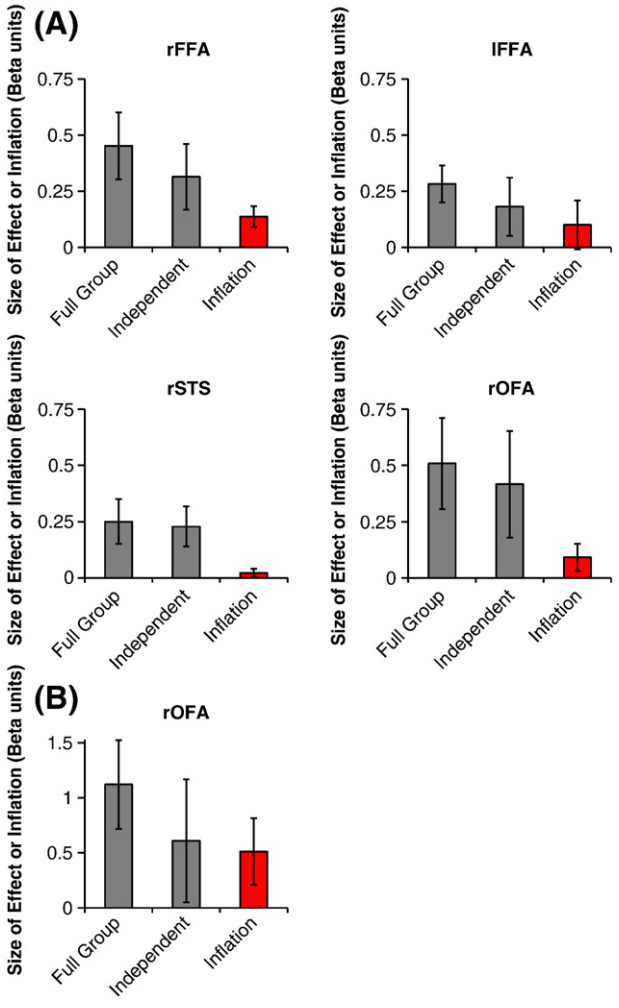

Fig. 1. Full group and LOSO effect sizes, as well as the inflation of effect sizes (full group minus LOSO; in red). Error bars represent 95% confidence intervals on the effect size across subjects (gray bars) and on the within-subject inflation (red bars). (A) The four regions identified in data set 1 (block design). (B) Right OFA in data set 2 (event-related design) replicates the finding from the block design in data set 1.

were some *a priori* rationale, could choose one of the two based on some consistent criteria (e.g., the most anterior cluster).

In data set 1, there were on average 6.4 clusters per LOSO GLM. Four were consistent across all LOSO GLMs: right superior temporal sulcus (STS), right middle fusiform gyrus (rFFA), right inferior occipital gyrus (rOFA), and left middle fusiform gyrus (lFFA). (These same four clusters were also present in the full group GLM.)

Effect size

Beta weights for each condition were extracted in each ROI for the full group GLM. The effect size was defined as the magnitude of the difference in beta weights that defined the ROIs:

$$\Delta\beta = \beta_{face} - [\beta_{house} + \beta_{scramb-face} + \beta_{scramb-house}] / 3.$$

This represented the potentially inflated estimate of face selectivity in these ROIs. Next, for each LOSO GLM, the beta weights for each condition were extracted from only the subject left out of the GLM that defined the ROI for that subject. Done iteratively, this yielded new beta weights for each subject, without bias due to non-independence. The effect size was then computed for the data extracted from the LOSO GLMs. Paired *t*-tests were conducted to compare the group/dependent effect size to the liberal independent and conservative independent estimates of effect size.

Data Set 2: event-related face/house discrimination task

Paradigm

In the event-related task, on each trial, an image of a house or a face was presented, starting from a random level of phase coherence and gradually cohering at a rate of 1% per 100 ms. The sequence ended at 74% coherence, because at this point all objects were clearly discriminable. Participants made either a gender discrimination (male/female) or house-size discrimination (one-story/two-story); they were instructed to respond as quickly as possible while minimizing errors. Using the right hand, button 1 was pressed for "Male" or "One-story", and button 2 was pressed for "Female" or "Two-story."

General linear models

Face-selective activation maps were defined by the contrast between face discriminations and house discriminations. Face and house regressors were defined as a canonical hemodynamic response function convolved with a stick function at the mean discrimination reaction time (68% phase coherence). A group random-effects GLM and nine LOSO GLMs were performed as described above for data set 1.

Regions of interest

One example region was defined in the group GLM to test the LOSO method in an event-related paradigm (Fig. 1B). This was a face-selective region in the right inferior occipital cortex. We chose this region in order to replicate the LOSO technique with the same functional region, while using a different data set and paradigm (data set 2 is event-related while data set 1 is a block design).

In each LOSO GLM, several clusters beyond the right OFA were identified. Again, for each of these clusters, we identified the anatomical neighborhood of these clusters. We then examined all LOSO ROIs to determine whether activity in a given anatomical neighborhood (e.g., gyrus or sulcus) was consistent across all GLMs. On average, 11.8 clusters were present per LOSO GLM. Six were consistent across all LOSO GLMs, including right OFA, right temporal pole, left temporal pole, precuneus, medial lingual/striate visual cortex, and anterior cingulate. (In the full group GLM, five of these six regions were present, with only the precuneus failing to survive cluster-correction.)

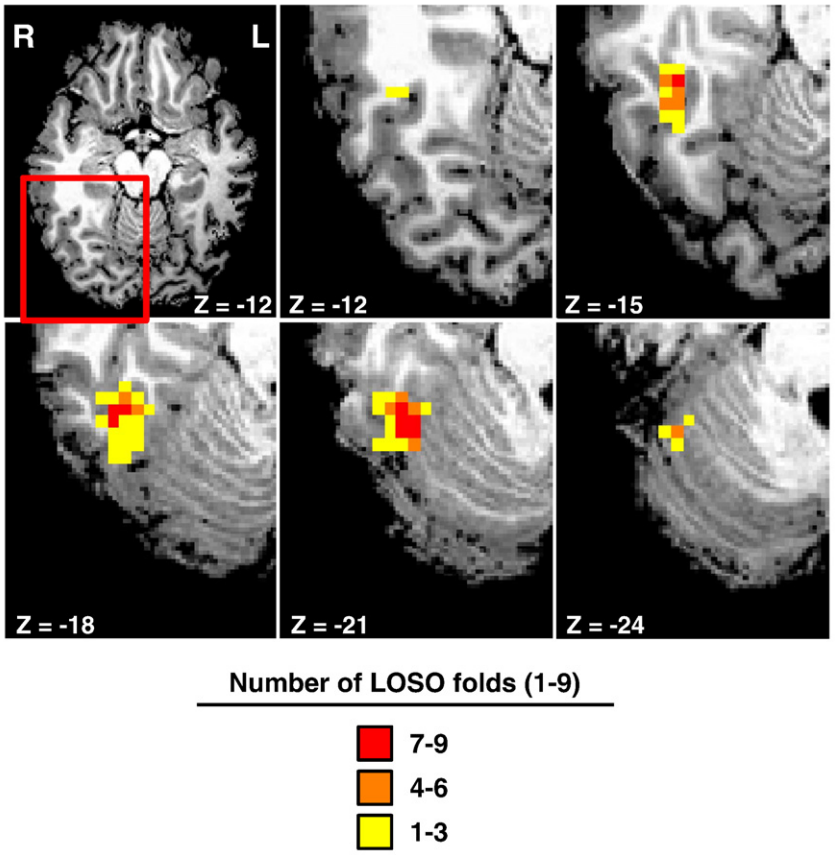

Fig. 2. The overlap of LOSO ROIs in an example region (right FFA). Voxels shaded yellow passed threshold in 1–3 of 9 LOSO GLMs; orange, 4–6 LOSO GLMs; red 7–9 LOSO GLMs.

Effect size

Beta values for each condition were extracted in the ROI for the group GLM. The effect size was defined as the magnitude of the difference in beta weights that defined the ROI: $\beta_{face}-\beta_{house}$. LOSO GLM effect sizes were calculated in the same way, but using the data from the left-out subject, as for data set 1. Paired *t*-tests were again conducted to compare the group/dependent effect size to the liberal independent and conservative independent estimates of effect size.

Results

In Fig. 2, we illustrate the LOSO ROIs, and the degree to which they overlap, within the right FFA. A significant number of voxels in the rFFA only survive a subset of the LOSO folds. However, some voxels in the right FFA survive in each LOSO fold.

Effect size measured in the group GLM-defined ROIs was significantly inflated relative to the corresponding conservative independent LOSO-defined ROI in 4 of 5 cases (Fig. 1A: rSTS, *t*-test $p<0.05$, 7/9 subjects inflated, sign test $p<0.09$; rFFA, *t*-test $p<0.001$, 9/9 subjects inflated, sign test $p<0.01$; lFFA, *t*-test $p<0.07$, 7/9 subjects inflated, sign test $p<0.09$; rOFA, *t*-test $p<0.01$, 8/9 subject inflated, sign test $p<0.05$. Fig. 1B: rOFA, *t*-test $p<0.01$, 8/9 subjects inflated, sign test $p<0.05$). Because the LOSO GLMs have less power (fewer degrees of freedom), we also defined the LOSO ROIs with a more liberal threshold (see Methods). The group GLM was also significantly inflated relative to these liberal LOSO-defined ROIs (p values<0.05 except lFFA [$p<0.07$], not shown). The conservative and liberal LOSO-defined ROI effect sizes differed in the rFFA (0.315 vs. 0.377, $p<0.05$), indicating that in this case, the more liberal threshold for the independent estimation of ROI led to a larger observed effect size (albeit smaller than in the group GLM). Otherwise, no significant differences were found between the conservative versus liberal LOSO threshold.

Discussion

The LOSO analyses presented here confirm that estimated effect sizes can be significantly inflated when using a (non-independent) whole-brain group analysis relative to an independent test after a leave-one-subject-out cross-validation procedure. The LOSO procedure reduces this inflation by defining ROIs with an independent data set.

It is worth pointing out that an inflated average effect size in a region is not necessarily problematic. For example, if an investigator wishes only to identify regions exhibiting significant activation for a given contrast, the LOSO technique is unnecessary because the magnitude of the effect size is not relevant to the investigator's claim. However, if an estimate of the effect size is desired, an independent test, such as the LOSO procedure, is necessary and altogether straightforward to carry out. It is true that statements of effect size within whole-brain-defined ROIs are "redundant" and "statistically guaranteed" (Vul and Kanwisher, in press), but they may be useful nonetheless in order to provide a more intuitive account of one's findings. Nevertheless, it is always necessary for readers and authors to be aware of the potential inflation that will be inherent in any reported effect when it is derived from an ROI that is based on a non-independent group whole-brain GLM analysis.

In addition to estimating effect sizes that are free of the non-independence error, LOSO cross-validation can be used to identify independent ROIs for further statistical inferences, such as time course extraction and analysis or further contrasts that are not orthogonal to the original contrast used to define the ROIs (e.g., A + B vs. C + D, followed by A vs. C).

One could also use LOSO to independently estimate correlations. One simple procedure is as follows: For each LOSO GLM, find a LOSO ROI with activity that is significantly correlated with a behavioral/personality measure across subjects. Activity (or beta weights) in these voxels is then computed for the subject left out. This is repeated for every LOSO GLM, so that each subject will have contributed an independent-ROI activity (or beta) value and a behavioral/personality measure. Behavioral measures and brain activity can subsequently be used to estimate the correlation, minimizing bias due to the non-independence error. Although this procedure will radically reduce inflation of effect size, simulations suggest that there remains the possibility of some residual inflation (Ed Vul, 2009, personal communication), especially when considering a small number of voxels.

An important application of LOSO is to identify ROIs for the left-out subject as an alternative to a separate, independent, functional localizer. This could be useful in a number of situations. Independent localizers often consist of a simple task (e.g., a 1- or 2-back working memory task) to define a functional region or regions for subsequent analysis in a critical task that is tailored to the specific hypothesis at issue in the study. To the degree that different brain areas are recruited under different cognitive demands, a localizer may fail to identify the voxels of greatest interest. A LOSO "localizer" consisting of the critical task itself maintains identical cognitive demands and independence (e.g. Esterman et al., 2009). In other cases, a separate independent localizer is simply not available. Finally, in many situations, the regions identified in the whole-brain analysis were not predicted *a priori*, and thus an independent localizer was not conducted in the original experiment.

In our examples, we included only neighborhoods that appeared in *all* LOSO GLMs. One could also explore regions that survive some subset of LOSO GLMs (e.g., 8 of 9, where this number could be based on a non-parametric sign test: 8/9, $p<0.02$). Regions can be examined if they survive any number of LOSO folds; however, if a region fails to survive the threshold in a LOSO fold, the subject-left-out should be eliminated from the analysis, incurring a loss of power. Another alternative is to use a more liberal threshold for the LOSO GLMs, if necessary, due to the decrease in degrees of freedom for each LOSO analysis (number of subjects-1). This is analogous to the case in which independent (within-subject) localizers are analyzed at very liberal statistical thresholds; this is appropriate because the localizer data are independent of the subsequent analysis, and thus any noise in a localizer is independent of the experimental manipulation in the main experiment. The use of a liberal threshold at intermediate stages of analysis (e.g., using a functional localizer to select voxels for use in a secondary analysis) in combination with an appropriately conservative threshold at the secondary/final analysis remains an appropriate choice. In practice, we suggest that experimenters only apply LOSO to an anatomical neighborhood in which they know an effect exists from the group analysis, but importantly, *without* the constraint that the LOSO regions necessarily overlap with the group ROI. This way, if a non-zero effect exists, it will survive all or most LOSO folds in the more generally defined region.

The group GLM ROI cannot be used to constrain the definition of LOSO ROIs, as that would be a form of "peeking" and create the possibility for non-independence. Instead, the extent of the anatomical neighborhood should be defined based on existing theory and previous literature, and only guided more generally by the group analysis. Plotting the LOSO ROIs as in Fig. 2, the degree to which LOSO ROIs overlap and the extent of the anatomical neighborhood can be transparently displayed. Nevertheless, one could argue that the anatomical neighborhood approach could still be biased. If the "anatomical neighborhood" approach is seen as introducing a form of bias (because the anatomical neighborhood – even if defined as a gyrus or Brodmann area – was identified using the result of the full group GLM) one could instead conservatively include all voxels from the entire brain that survived each LOSO fold to define the left-out subject's ROI, without regard to anatomical location. The voxels that survive each fold could be plotted as in Fig. 2 to be transparent

regarding the similarity of the ROIs. The experimenter can do as he or she sees fit with a full understanding of the consequences of either choice.

Cross-validation is not a novel idea, but as far we know, its application in the fMRI community has been limited to within-subject multivoxel pattern classification (i.e., leave-one-run-out) analyses and has not been extended to random-effects (subject as the factor) GLMs in this way. The LOSO method for ROI definition is computationally and conceptually simple and allows for independent estimates of effects (r, t, z, etc.) initially identified from group-level whole-brain contrasts as well as permitting secondary analyses (e.g., multivoxel pattern classification or effect size estimates) that are free of the non-independence error. The method described here provides a tool to avoid some of the potential pitfalls highlighted in the recent literature.

Acknowledgments

We would like to thank Amy Shelton, Ed Vul, and an anonymous reviewer for useful comments and feedback. This research was supported by NIH grant R01-DA13165 (S.Y.).

References

Allison, T., Ginter, H., McCarthy, G., Nobre, A.C., Puce, A., Luby, M., Spencer, D.D., 1994. Face recognition in human extrastriate cortex. J. Neurophysiol. 71, 821–825.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. 16, 4207–4221.

Esterman, M., Yantis, S., 2009. Perceptual expectation evokes category-selective cortical activity. Cerebral Cortex.

Esterman, M., Chiu, Y.C., Tamber-Rosenau, B.J., Yantis, S., 2009. Decoding cognitive control in human parietal cortex. Proceedings of the National Academy of Sciences 106, 17974–17979.

Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C., Anderson, A.W., 2000. The fusiform "face area" is part of a network that processes faces at the individual level. J. Cogn. Neurosci. 12, 495–504.

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., Martin, A., 1999. The effect of face inversion on activity in human neural systems for face and object perception. Neuron 22, 189–199.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–540.

Lieberman, M., Berkman, E., Wager, T., 2009. Correlations in social neuroscience aren't voodoo: a reply to Vul et al. Perspect. Psychol. Sci. 4.

Lindquist, M.A., Gelman, A., 2009. Correlations and multiple comparisons in functional imaging: a statistical perspective (Commentary on Vul et al., 2009). Perspect. Psychol. Sci. 4.

Nichols, T.E., Poline, J.B., 2009. Commentary on Vul et al.'s (2009) "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". Perspect. Psychol. Sci. 4.

Puce, A., Allison, T., Gore, J.C., McCarthy, G., 1995. Face-sensitive regions in human extrastriate cortex studied by functional MRI. J. Neurophysiol. 74, 1192–1199.

Puce, A., Allison, T., Bentin, S., Gore, J.C., McCarthy, G., 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18, 2188–2199.

Rossion, B., Caldara, R., Seghier, M., Schuller, A.M., Lazeyras, F., Mayer, E., 2003. A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing. Brain 126, 2381–2395.

Sergent, J., Ohta, S., MacDonald, B., 1992. Functional neuroanatomy of face and object processing. A positron emission tomography study. Brain 115 (Pt. 1), 15–36.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4.

Vul, E., Kanwisher, N., in press. Begging the question: The non-independence error in fMRI data analysis. In: Foundations and Philosophy for Neuroimaging (Hanson S, Bunzl M, eds).