Variability in FMRI: A Re-Examination of Intersession Differences

FMRIB Technical Report TR04SS1
(A related paper has been accepted by Human Brain Mapping)

**S.M. Smith, C.F. Beckmann, N. Ramnani, M.W. Woolrich,
P.R. Bannister, M. Jenkinson, P.M. Matthews, D.J. McGonigle**

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK

(DM is at Laboratoire de Neurosciences Cognitives et Imagerie
Cérébrale, Hôpital de la Salpêtrière, CNRS UPR 640 - LENA, Paris)

Abstract

We revisit the McGonigle et al (2000) paper on intersession variability, showing that, contrary to one popular interpretation of the original paper, intersession variability is not necessarily high. We also highlight how evaluating variability on the basis of thresholded single-session images alone can be misleading. Finally, we show that the use of different first-level pre-processing, time-series statistics and registration analysis methodologies can give significantly different intersession analysis results.

Keywords: FMRI, session variability, reproducibility, longitudinal studies

1 Introduction

The magnitude of the BOLD (Blood Oxygenation Level Dependent) effect in FMRI, a marker of neuronal activation, is often only of similar magnitude to the noise present in the measured signal. In order to increase power and to allow conclusions to be made about subject populations, it is common practice to combine data from multiple subjects. It is also common to take multiple sessions from each subject, again to increase sensitivity to activation, or for other experimental design reasons, such as tracking changes in function over time. Therefore it is important that inter-session variability present in FMRI data be understood, and in response, McGonigle et al. [18] presented an in-depth study of this issue.

In designing both multi-subject and single-subject multi-session studies, it is critical for the experimenter to have some idea of the relative sizes of within-session variance and inter-session variance. For example, if inter-session variance is large, it could be difficult to detect longitudinal experimental effects (e.g., in studies of learning [21] and post-stroke recovery [15]). If FMRI is to be used in pre-surgical mapping (e.g., [9]), which, by its nature will involve only a single subject, correct interpretation will be dependent on an appreciation of the potential uncertainty due simply to a session effect. In multi-subject studies, it is advantageous to have some idea of the expected inter-session variance, as this will contribute to the observed inter-subject variance.

In order to investigate how well a single session dataset from a single subject typified the subject's responses across multiple sessions, McGonigle and colleagues [18] carried out the same FMRI protocol on 33 separate days; on each day 3 paradigms were run (visual, motor and cognitive), and the variation in "activation" was studied. This paper drew three main conclusions: i) the use of "voxel-counting" on thresholded statistical maps was not an ideal way to examine reproducibility in FMRI; ii) a "reasonably large" number of repeated sessions was essential to properly estimate inter-session variability, and iii) the results of a single session on a single subject should be treated with care if nothing was known about inter-session variability.

While [18] noted the presence of between-session variability in their experiment, they did not attempt to systematically assess the causes of this variance. There are a number of potential contributors; physiological variance (subject), acquisition variance (scanner), and also differences in analysis methodology and implementation. As noted in the original paper, "it is possible that spatial preprocessing (for example) may affect inter-session variance quite independently of underlying physical or physiological variability". This view is supported by [19], where analysis methodology is shown to affect apparent intersession variance. Here we revisit the analysis of McGonigle's data and consider session variability in the light of the effects that different first-level processing methods can have.

Furthermore, some readers (e.g., [6, 4]) have taken from [18] the simple broad-brush conclusion that there was a "large amount of session variability". One of the purposes of this paper is to address this misconception; for example, Section 4.3 shows that in fact inter-session variability was of similar magnitude to within-session variability in this dataset.

We start with a brief theoretical overview of the components of variance present in multiple-session data. We then describe the original data and analysis, as well as the new analyses carried out for this study, with explanation of the measures used in this study to assess session variability. We then present the variability results as found from this data, centring around the use of mixed-effects Z values in relevant voxels as the primary measure of interest. We also show qualitatively why it is dangerous to judge variability through the use of thresholded single-session images.

2 Variance Components

Researchers often refer to different "group analyses", the most common being "fixed-effects" and "mixed-effects". What these terms are actually referring to are different inter-session (or inter-subject) noise (variance) models. We now summarise what the terms and associated models mean.

We start with the equation for the t statistic:

$$t = \frac{\text{mean effect}}{\sqrt{\text{variance(mean effect)}}}, \tag{1}$$

i.e., we are asking how big the mean effect size is compared with the "noise" (the mean's standard deviation[1]). The standard deviation is the square root of either the fixed-effects variance of the mean or the random-effects variance of the mean.

With fixed-effects modelling, we assume that we are only interested in the factors and levels present in the study, and therefore our higher-level fixed-effects variance FV is derived from pooling[2] the first-level (within-session) variances (of first-level effect size mean) FV_i, according to:

$$FV = \frac{\sum(FV_i)}{n^2}, DoF_{FV} = \sum DoF_{FVi}, \tag{2}$$

where DoF is the degrees of freedom, which, in the case of FMRI time series, is usually large. This modelling therefore ignores the cross-session (or cross-subject) variance completely and the results cannot be generalised outside of the group of sessions/subjects involved in the study.

With simple mixed-effects[3] modelling, we derive the mixed-effects variance MV directly from the variance of the first-level parameter estimates PE_i (effect sizes) or contrasts of parameter estimates:

$$MV = \frac{var(PE_i)}{n}, DoF_{MV} = n - 1, \tag{3}$$

with a (normally) much smaller DoF than with fixed-effects. Thus the modelling uses the cross-session (or cross-subject) variance, and the results (which are generally "more conservative" than with a fixed-effects analysis) are relevant to the whole population from which the group of sessions/subjects was taken.

The mixed-effects variance is the sum of the fixed-effects (within-session) variance and random-effects (pure inter-session) variance (though note that simple estimation methods calculate this directly, as above, and do not explicitly use the fixed-effects variance). Therefore the estimated mixed-effects variance should in theory and in practice be larger than the fixed-effects variance. We expect that when there is large inter-session variance there will be a large difference between fixed- and random-effects analyses.

There have recently been significant developments in group-level analysis. For example, it has been shown in [1] that there is value in carrying up lower-level variances to higher-level analyses of mixed-effects variance, and one implementation of this, using Bayesian modelling/estimation methodology has been reported in [3]. Whilst the dataset used in this paper may well prove useful in investigating these developments further, this is beyond the scope of this paper. Instead, this paper concentrates primarily on two other questions, namely the magnitude of session variability, and the effect that *first-level* analysis methodologies can have on its effect. Therefore, for mixed-effects analyses in this paper, we have only used ordinary least squares (OLS) estimators (see equation 3 and [12]).

[1] Note that in the simplest cases the variance of the mean is the variance of the residuals divided by the number of data points.

[2] The first factor of $1/n$ in FV comes from taking the mean of the first-level variances, i.e., pooling them, and the second factor comes from converting this higher level variance from a variance of residuals into the variance of the (higher-level) mean. For more detail see [16].

[3] Note that the terms "mixed effects" and "random effects" are often (incorrectly) used interchangeably.

3 Methods

3.1 Original Experiments and Analysis

We now describe the experiment and original analysis carried out in [18]. A healthy 23-year-old right-handed male was scanned on 33 separate days (over 2 months) with as many factors as possible held constant. On each day, three block-design paradigms were run (all using rest=24.6s, activation=24.6s block lengths): visual (8Hz reversing black-white chequerboard, 36 time points after deleting the first two), motor (finger tapping, right index finger at 1.5Hz, 78 time points) and cognitive (0.66Hz random number generating vs counting, 78 time points), with the paradigm order randomised. The data was collected on a Siemens Vision at 2T; TR=4.1s, 64x64x48 3x3x3mm voxels. A single T1-weighted 1.5x1x1mm structural scan was taken.

Original analysis was carried out with SPM99 (www.fil.ion.ucl.ac.uk/spm). All 99 sessions were realigned (motion-corrected) to the same target (the first scan of the first session of the first day) and then a mean over all 99 sessions was created. This was used to find normalisation (to a T2-weighted target in MNI space [8]) parameters for all 99 sessions (using 12-parameter affine followed by 7x8x7 basis-function nonlinear registration). Sinc interpolation on final output was used.

Sessions containing "obvious movement artefacts" were identified by eye and removed from consideration (3 motor, 2 visual, 3 cognitive). Cross-session analysis was carried out for voxels in standard space which were present in all sessions. Spatial filtering with a Gaussian kernel of FWHM 6mm was applied. Each volume of each session was intensity-normalised (rescaled) so that all had the same mean intensity.

Voxel time-series analysis was carried out using general linear modelling (GLM). The data was first pre-coloured by temporally smoothing the data with a Gaussian of 6s FWHM. Slow drifts in the data were removed by including drift terms in the model - a set of cosine basis functions effectively removing signals of period longer than 96s.

For presentation of within-session results, voxel-wise thresholding ($p < 0.05$) was used, correcting for multiple comparisons using Gaussian random field-theory (GRF) [11].

Both fixed- and "random"-effects analyses were carried out to examine the effects of using different variance components, and an extra-sum-of-squares (ESS) F-test was performed across all sessions of each paradigm to assess the presence of significant inter-session variance.

3.2 Methods Tested

We now describe the analysis approaches used for this paper. The two packages used for our investigations were SPM99b (Statistical Parametric Mapping, www.fil.ion.ucl.ac.uk/spm) and FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl, version 1.3, June 2001). Both are freely available and widely used.

SPM includes a motion-correction ("realignment") tool, a tool for registration ("normalisation") to standard-space, GLM-based time-series statistics [23] and GRF-based inference [11]. SPM carries out standard-space registration before time-series statistics. The SPM99b time series statistics correct for temporal smoothness by precolouring [10].

GLM-based analysis in FSL is carried out with FEAT (FMRI Expert Analysis Tool), which uses other FSL tools such as BET (Brain Extraction Tool [20]), an affine registration tool (FLIRT - FMRIB's Linear Image Registration Tool [14, 13]), and a motion-correction tool based on FLIRT (MCFLIRT [13]). FEAT carries out standard-space registration after time-series statistics. FSL time series statistics correct for temporal smoothness by applying prewhitening, as described in [22].

6 different, complete analyses were carried out with various combinations of pre-processing and time-series statistics options, in order to allow a variety of comparisons to be made. In tests A,C,G, FSL was used for preprocessing and registration whilst in tests D,E,F, SPM was used. For tests A,D,G, FEAT time-series statistics was used whilst for C,D,F, SPM time-series statistics was used.

In tests A-E the various controlling parameters were kept as similar as possible, both to each other and to default settings in the relevant software packages. Tests A vs D and C vs E hold the statistics method constant whilst comparing spatial methods, therefore showing the relative merits of the "spatial" components (motion correction and registration). Tests A vs C and D vs E hold the spatial method constant whilst comparing statistical components, thus showing the relative merits of the statistical components (time-series analysis). A vs E tests pure-FSL against pure-SPM. F and G test pure-SPM and pure-FSL respectively, with these analyses set up to match the specifications of the original analyses in [18] as closely as possible, including turning on intensity normalisation in both cases. For a summary, see Table 1.

(For B, ICA-based temporal-model-free analysis was carried out; the model-free results are not included in this paper, but will be presented elsewhere.)

Because the methods for high-pass temporal filtering in FSL and SPM are intrinsically different, they cannot be set to act in exactly the same way (within A-E and within F,G) by choosing the same cutoff period in each; instead, the cutoff choices were made so as to match as closely as possible the extent to which the relevant signal and noise frequencies were attenuated by the different methods. For the purposes of this paper, high-pass temporal filtering is considered to be part of the temporal statistics, where it most naturally fits.

The non-default "Adjust for sampling errors" motion-correction option in SPM was not used.

8 sessions (of the 99) were excluded from the original analysis in [18], due to "obvious movement artefacts". These were however included in our analyses as we did not consider that there was sufficient objective reason to exclude them; the estimated motions for these sessions were not in general significant outliers relative to the average motion across sessions and any apparent (activation map) motion artefacts were not in general significantly different from the majority of the sessions. The quantitative results given in Sections 4.3 and 4.6 were in fact recalculated without these 8 sessions (i.e., reproducing the same dataset as used in [18]), but without any significant change in results, and therefore are not reported here.

3.3 Inter-Session Evaluation Methods

For all paradigms and all analysis methods, simple fixed-effects (FE) and OLS mixed-effects (ME) Z-statistics were formed. For each paradigm a mask of voxels which FE considered potentially activated ($Z > 2.3$) was created. This contains voxels in which a ME analysis is potentially interested (given that ME generally gives lower Z-statistics than FE[4]). This mask was averaged over A,C,D and E to balance across the various methods, and then eroded slightly (2mm in 3D) to avoid possible problems due to different brain mask effects.

We initially investigated the size of intersession variance, by estimating the ratio of random effects variance to fixed

[4]We are attempting to identify voxels of potential interest in ME-Z images; given that ME-Z can be thought of as being related to FE-Z but scaled down by a factor related to session variance, this seems like a good way of choosing voxels which have the potential to be activated in the ME-Z image, depending on the session variance. In order to investigate the dependency of this approach on the FE-Z threshold chosen, we re-ran the tests leading to the ME-Z plots presented in Figure 8, having determined the regions of interest using a lower FE-Z threshold (Z>1.64, i.e., a factor of 5 more liberal in the significance level). The mean ME-Z results were all scaled down, as expected, but the qualitative (i.e., relative) results were exactly identical to those presented in Figure 8.

	Preprocessing	Statistics	Registration
A	FSL (MCFLIRT spat=5 intnorm=n)	FSL (FEAT) (hp-FSL=40)	FSL (FLIRT)
C	FSL (MCFLIRT spat=5 intnorm=n)	SPM (hp-cos=72)	FSL (FLIRT)
D	SPM (SPM-mc&norm spat=5 intnorm=n)	FSL (FEAT) (hp-FSL=40)	SPM (done in preproc)
E	SPM (SPM-mc&norm spat=5 intnorm=n)	SPM (hp-cos=72)	SPM (done in preproc)
F	SPM match [18] (SPM-mc&norm spat=6 intnorm=y)	SPM (hp-cos=98.4)	SPM (done in preproc)
G	FSL match [18] (MCFLIRT spat=6 intnorm=y)	FSL (FEAT) (hp-FSL=53)	FSL (FLIRT)

spat - spatial filtering with full-width-half-maximum given in mm.
intnorm - intensity normalisation - the intensity rescaling of each volume in a 4D FMRI data set so that all have the same mean within-brain intensity.
hp-FSL - FSL's high-pass temporal filtering with cutoff period given in seconds.
hp-cos - high-pass temporal filtering (in seconds) via cosine basis functions.

Table 1: Different analyses carried out.

effects, averaged over the voxels of interest as defined above. Given that ME variance is the sum of FE and RE variance, we estimated the RE (intersession) variance by subtracting the FE variance from the ME variance. We then took the ratio image of RE to FE variance, and averaged over the masks described above. This ratio would be 0 if there was no intersession variability and rises as its contribution increases. A ratio of 1 occurs when intersession and intrasession variabilities make similar contributions to the overall measured ME variance.

Next, we investigated whether session variability is indeed Gaussian distributed. If it is not, then inference based on the OLS method used for ME modelling and estimation in this paper would need much more complicated interpretation (as also would be the case with many other group-level methods used in the field). We used the Lilliefors modification of the Kolmogorov-Smirnov test [17] to measure in what fraction of voxels the session effect was significantly non-Gaussian.

The variance ratio figures do not take into account estimated effect size, which in general will vary between methods, and so the primary quantification in this paper uses the mixed-effects Z (ME-Z). This is roughly prortional to the mean effect size and inversely proportional to the intersession variability. This makes ME-Z a good measure with which to evaluate session variability; it is directly affected by the variability, while being weighted higher for voxels of greater interest (i.e., voxels containing activation). We are not particularly interested in variability in voxels which contain no mean effect. We therefore base our cross-subject quantitations on ME-Z comparisons within regions of interest (defined above).[5]

If one of the analysis methods tested here results in increased ME-Z, then this implies reduced overall method-related error (increased accuracy) in the method. This is due to the fact that unrelated variances add; whilst a single-session analysis cannot eliminate true inter-session variance instrinsic to the data, it can add ("induce") variance to the effective intersession variance due to failings in the method itself (for example, poor estimates of first-level effect/variance, or registration inaccuracies). Therefore the best methods should give ME-variance

[5]Whilst we are primarily investigating analysis efficiency and session variance by looking at regions of potential activation, note that it is also necessary to ensure that the non-activation (null) part of the ME distribution is valid, i.e., not producing "incorrect" numbers of false-positives; this investigation/correction of the ME null distribution is addressed below and uses the whole ME-Z image, not just the regions of potential activation.

which approaches (from above) the true, intrinsic inter-session ME-variance. (Remember that the same simple OLS second-level estimation method was used for all analyses carried out - it is only the first-level processing that is varied.)

Mean ME-Z was then calculated within the FE-derived masks. However, as well as reporting these “uncorrected” mean ME-Z values, we also report the mean values after adjusting the ME-Z images for the fact that in their histograms (supposedly a combination of a null and an activation distribution) the null part, ideally a zero mean and unit standard deviation Gaussian, was often significantly shifted away from having the null peak at zero. This makes Z values incomparable across methods, and needed to be corrected for. The causes of this effect include (spatially) structured noise in the data and in differences in the succcess between the different methods for correcting for temporal smoothness (a problem enhanced potentially for all methods given the unusually low number of time points in the paradigms).

We used two methods to correct ME-Z for null-distribution imperfections, and report results for both methods. With *hand-corrected peak shift correction*, the peak of the ME-Z distribution was identified by eye and assumed to be the mean of the null disribution; the ME-Z image then had this value subtracted. With *mixture-model-based null shift correction*, a (non-spatial) histogram mixture model was automatically fitted to the data using expectation-maximisation. This involved a Gaussian for the null part, and gammas for the activation and deactivation parts [2]. The centre of the Gaussian fit was then used to correct the ME-Z image. The advantage of the hand-corrected method is that it is potentially less sensitive to failings in the assumed form of the mixture components; the advantage of the mixture-model-corrected fit is that it is fully automated and therefore more objective.

It is not yet standard practice (with either SPM or FSL) to correct for null-Z shifts in ME-Z histograms; the most common method of inference is to use simple null-hypothesis testing on uncorrected T or Z maps (typically via Gaussian random field theory). By correcting for the shifts, what we are able to investigate the effects of using the different individual analysis components in the absence of confounding effects of null distribution imperfections.

Figure 1 shows an example ME-Z histogram including the estimates (by eye and by mixture-modelling) of the null mode. The estimated ME-Z shifts which were applied to the mean ME-Z values before comparing methods are plotted for all analysis methods and all paradigms in Figure 2. The shift is clearly more related to the choice of time-series statistics method than the choice of spatial processing method (motion correction and registration), but there is no clear indication of one statistics method giving greater shift extent than another. The two correction methods are largely in agreement with each other.

4 Results and Discussion

4.1 Fixed-Effects Activation Maps

The FE-based mask images (used to define the voxels used in the quantitative analyses reported below) are shown in Figure 3, overlaid onto the MNI152 standard head image.

4.2 Intersession Effect Size Plots

For analysis methods A and E we now show the effect size and its (fixed effects, within-session) temporal standard deviation, as a function of session number. Both the effect size and the temporal standard deviation are estimated as means over interesting voxels, as defined above. The plots were normalised by estimating the mean effect size over

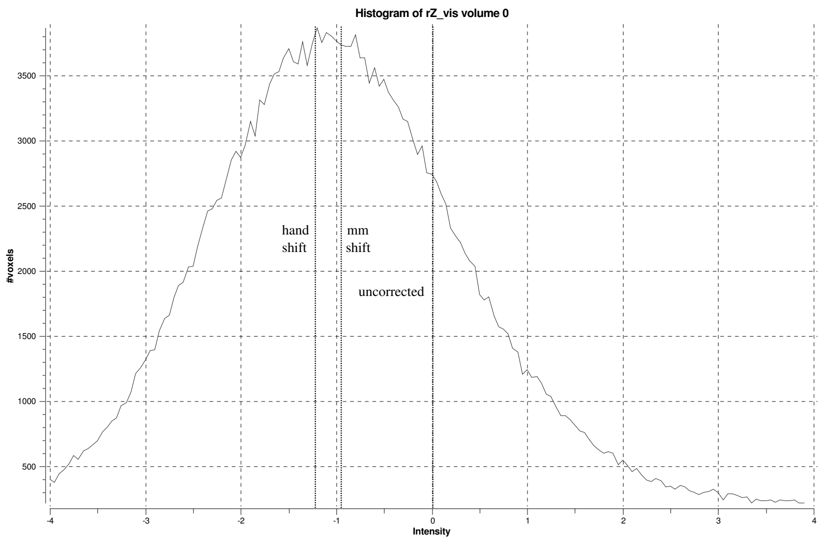

Figure 1: Example ME-Z histogram showing null-distribution shift, from analysis A of the visual paradigm.

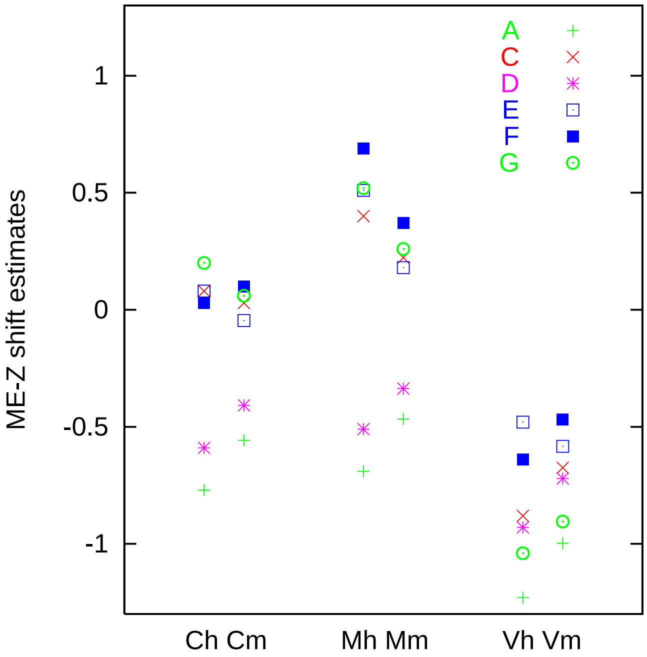

Figure 2: Estimated ME-Z null-distribution shifts. Different tasks: C=cognitive M=motor V=visual. Different correction methods: h=hand-shifted m=mixture-model-shifted.

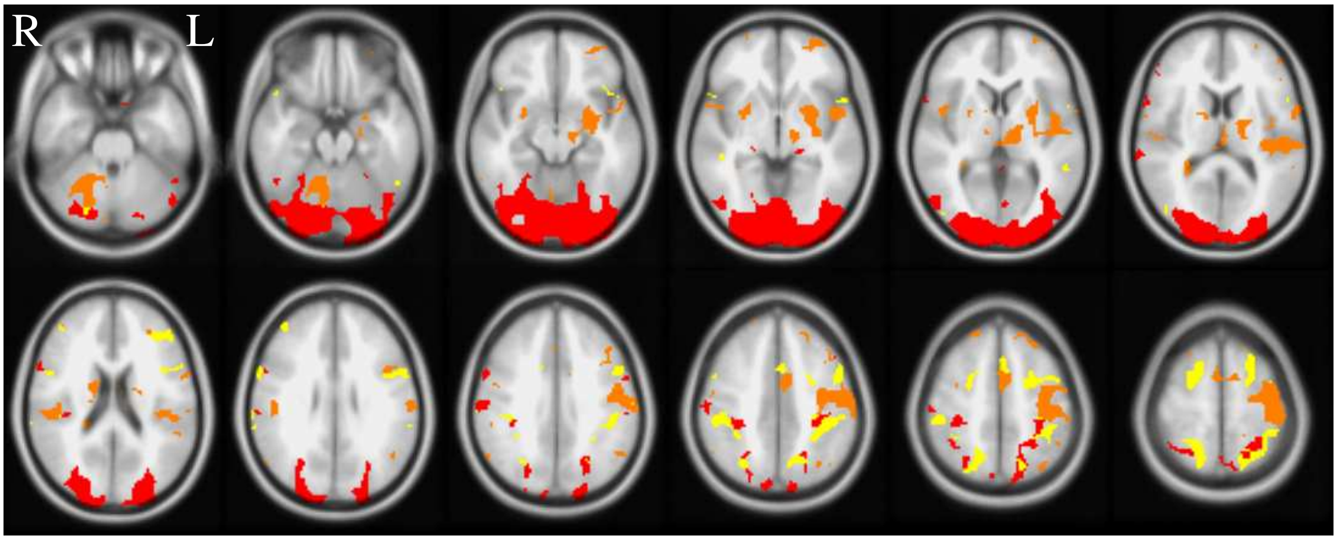

Figure 3: Masks of potentially activated voxels, within which mean ME-Z was calculated for each analysis method; red=visual, orange=motor, yellow=cognitive.

all sessions, scaling this to be unity, scaling the standard deviation by the same factor, and demeaning the effect size plot. See Figures 4 and 5. These plots show (as does the following section) that the within-session variance has similar magnitude to the inter-session variance. They also show that variability in effect size is higher than variability in its standard deviation (though the implication of this fact is not necessarily important to the primary points of this paper). Note that the results presented here correspond to the "uncorrected" plot in Figure 8.

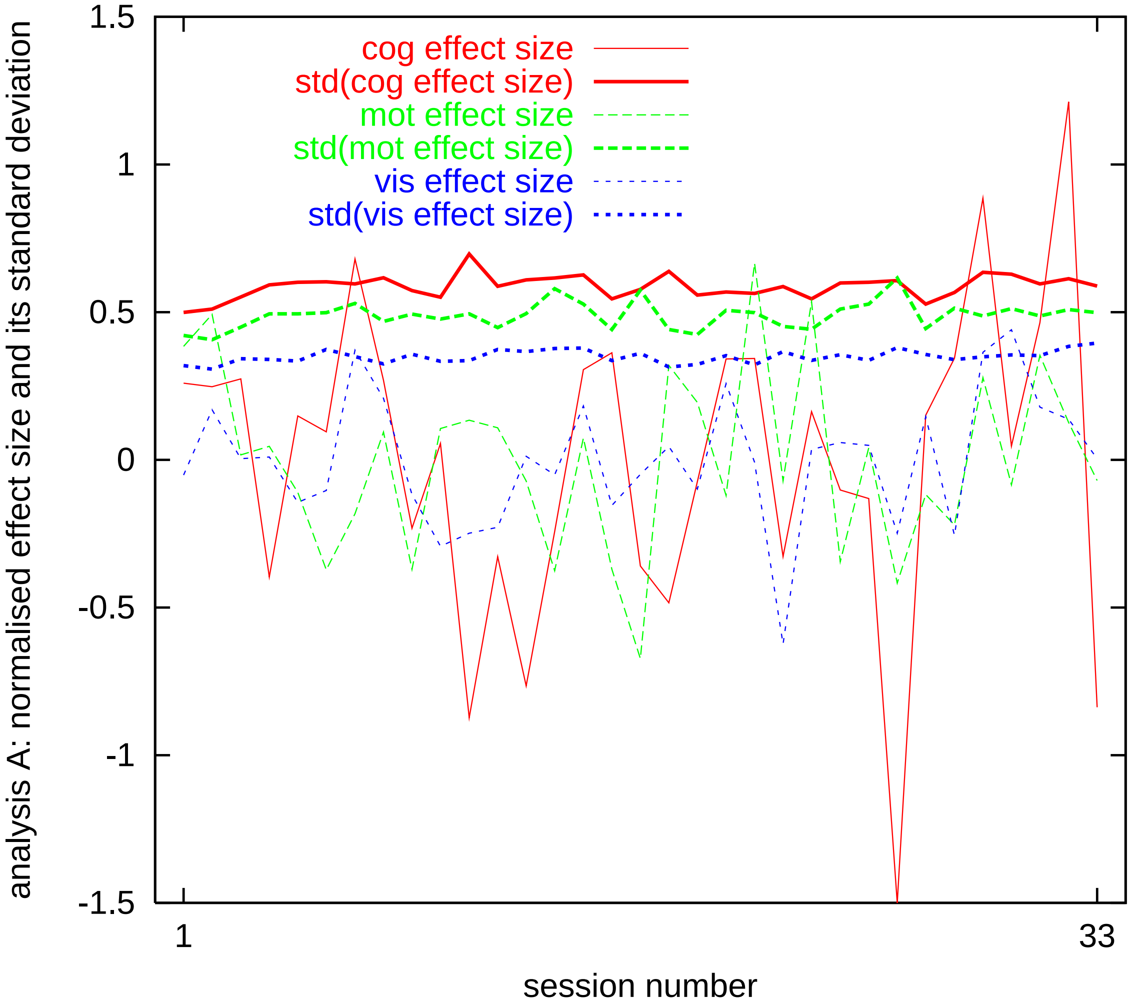

Figure 4: Mean first-level effect size and its (within-session) standard deviation, as a function of session number, for analysis A.

4.3 Quantification of Intersession Variance

In order to better quantitate the size of intersession variance, we estimated the mean ratio of RE (ME minus FE) to FE variance. Note here that any comparison between the RE and FE variance will be dependent on the number of time points in each session, with a larger number of time points leading to an increase in the RE:FE ratio.

The results are shown in Table 2. The interpretation of this is simple yet important; in these datasets, intersession variability is not large compared with within-session variability.[6]

[6]Noting the much greater variability (across methods) in these ratios than in the plots in figure 8, and by looking in detail at separate ME

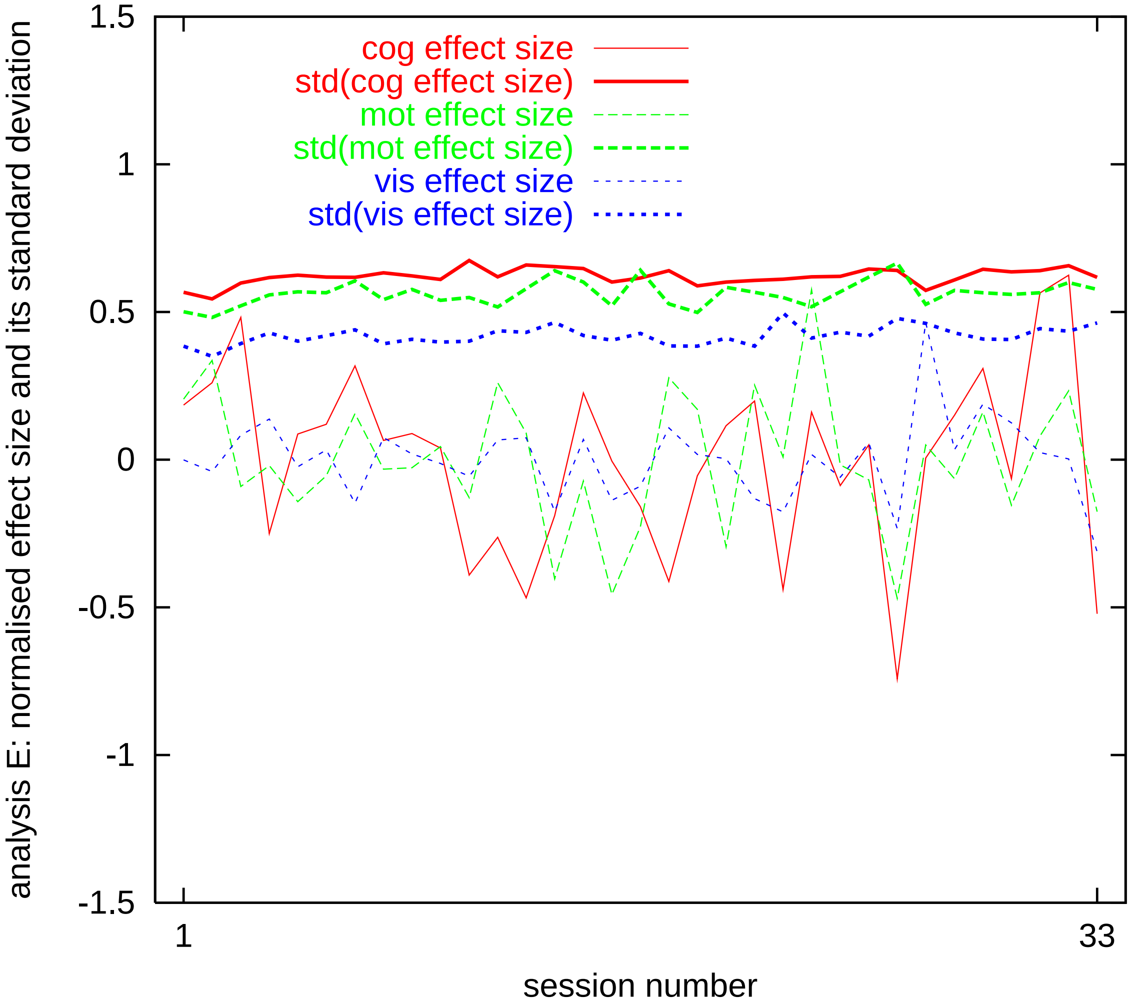

Figure 5: Mean first-level effect size and its (within-session) standard deviation, as a function of session number, for analysis E.

Note that we cannot make very useful interpretations of the variations across methods of the variance ratio, particularly without also taking into account the estimated effect size; hence the use of mixed-effects Z for the main method comparison results shown below in Section 4.6.

	A	C	D	E	F	G
cognitive	1.0	0.3	1.5	0.5	0.6	1.0
motor	0.9	0.3	1.4	0.6	0.6	1.1
visual	1.4	0.3	1.9	0.8	0.8	1.7

Table 2: Mean estimated ratio of RE (intersession) variance to FE (within session pooled) variance.

4.4 Test for Gaussianity of Intersession Variability

Using the results of analysis A, for each paradigm, we tested whether the session variability was Gaussian. At each voxel in standard space we took the (first-level) parameter estimates (effect sizes) from the relevant voxel in each of the 33 relevant first-level analyses, (i.e., the same data that was fed into the group-level ME analysis). The variance of these is the ME variance. For each set of 33 first-level parameter estimates, we ran the Lilliefors modification of the Kolmogorov-Smirnov test [17] for non-Gaussianity, with a significance threshold of 0.05. We would therefore expect, in null data, rejection of the Gaussianity null hypothesis at this 5% rate by random chance.

We calculated the fraction of voxels failing the normality test both across the whole brain, and within the FE-derived masks described above. In both cases, and for all three paradigms, the fraction of failed tests was less than 7.5% (range 4.5-7.3%), which is very close to the expected 5% rate of null-hypothesis rejections if in fact all the data is normal. This provides strong quantitative evidence for the normality of the session variability in this data. Qualitatively, the voxels where the null hypothesis was rejected were scattered randomly through the images, not clumped, again suggesting that they were rejected by pure random chance rather than because of some true underlying non-Gaussian process.

4.5 On (Not) Drawing Conclusions About Session Variability on the Basis of Thresholded Single-Session Images

[18] does not include any such statement as "session variability is high", or even any quantification explicitly suggesting in a simple way that session variability is a serious problem; nevertheless, unfortunately, many researchers (e.g., [6, 4]) seem to have taken these messages from the paper. One of the causes of this is the apparent variability in Figures 2-4 (in [18]), which show, for each paradigm, each session's thresholded activation image (as a single sagittal slice maximum intensity projection). All three figures give the impression of large intersession variability, even for the strong visual paradigm.

The most important point to make with respect to this issue is that it is not safe to judge intersession variability by looking at variability in thresholded statistic images. It is perfectly possible for two unthresholded activation images to not be statistically significantly different and yet one contain activation just over threshold and the other just under, giving the false impression of large variability. The fact that thresholds are in any case chosen arbitrarily increases the weakness of this method of judging variability.

---

and FE variances, it is clear that the variation in these figures across methods is primarily due to variation in FE variance. This is possibly caused by differences in the methods of correcting for temporal autocorrelation at first level.

To illustrate these issues, Figures 6 and 7 show single-session thresholded images from analysis F of the visual experiment. Figure 6 is created using the same threshold as in [18], namely p<0.05, corrected for multiple comparisons using Gaussian random field theory [11]. In contrast, Figure 7 is created using a reduced threshold (the t threshold used in Figures 6 is reduced by 33%). Obviously there is more apparent activation when the threshold is reduced (though, note that it has clearly not been reduced so far that there is generally a huge amount of spatially variable "noise" activation caused by this). However, the interesting point is that the subjective impression of intersession variability is much reduced.

Finally, a question arises as to why Figure 6, which should match the original figure in [18] (having been processed in the same manner) appears to show less variability than the original figures. This was found to be due to the fact that suboptimal timing was used in the original model generation (caused by a particular default setting of the point within a TR that the model is sampled, which also corresponds to the point during a TR when that time point's whole FMRI volume is assumed to have been instantaneously sampled; this default was changed between SPM99 and SPM99b). The re-analysis was more efficient at estimating activation as better-matched models were used, causing less apparent inter-session variability. As part of the investigation of this effect, we tested the variability in peak Z values as the model timing was changed slightly. The *mean across sessions(max across space(Z))* value for 5 different phase shifts of the model (-1 TR to +1 TR) were found to be 6.6, 7.5, 7.9, 7.5, 6.9 (model timing running from earlier to later respectively). This is quite a large effect for these phase shifts, given that the paradigm is a block design. This is another illustration of the danger of judging variability solely on the basis of thresholded results.

4.6 Mean ME-Z Plots

Mean ME-Z plots are shown in Figure 8. Higher ME-Z implies less analysis-induced inter-session variance, or, viewed another way, greater robustness to session effects.

Before discussing these plots it is instructive to get a feeling for what constitutes "significant" difference in the plots. Suppose that two ME-Z maps were, in these figures, separated by a Z difference of 0.25. This would correspond to a general relative scaling between the two maps of approximately $.25/6 = 4\%$. We are interested in the effect that this difference has on the final thresholded activation map. Therefore we can estimate this effect by thresholding a ME-Z map at a standard level and also at this level scaled by 4%. Thresholding at $P < 0.05$, when corrected (using Gaussian random field theory) for multiple comparisons, corresponds to a Z threshold of approximately 5. We therefore thresholded the three ME-Z images from analysis F at levels of $Z > 5$ and $Z > 5.2$. For the cognitive, motor and visual ME-Z images, this resulted in reductions in supra-threshold voxel counts by 11%, 8% and 6% respectively. These are not small percentages; we conclude that a difference in 0.25 between the various plots can be considered to be "significant" in terms of the effect on the final reported mixed-effects activation maps. (Note that these different thresholdings were carried out with two threshold levels *on the same ME-Z image for each comparison*, hence the previous criticism of not comparing thresholded maps is not relevant here.)

We now consider plots A,C,D,E, the various tests which attempted to match all settings both to each other and to default usage. Firstly, consider comparisons which show the relative merits of the "spatial" components (motion correction and registration); A vs D and C vs E hold the statistics method constant whilst comparing spatial methods. Next, consider comparisons which show the relative merits of the statistical components (time-series analysis); A vs C and D vs E hold the spatial method constant whilst comparing statistical components. Finally, A vs E tests pure-FSL against pure-SPM.

Plots F and G test pure-SPM and pure-FSL respectively, with these analyses set up to match the specifications of

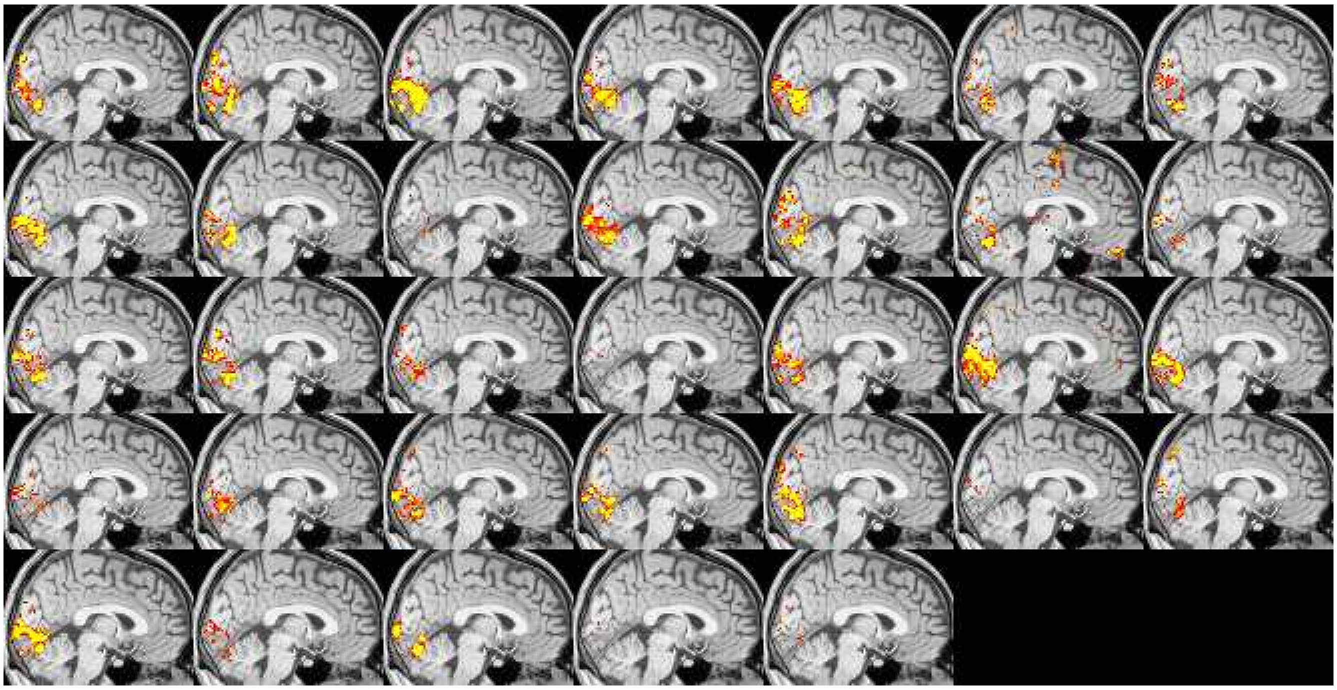

Figure 6: Visual paradigm; analysis F single-session thresholded maximum intensity projections, p<0.05 GRF-corrected. Each image corresponds to a different day's dataset.

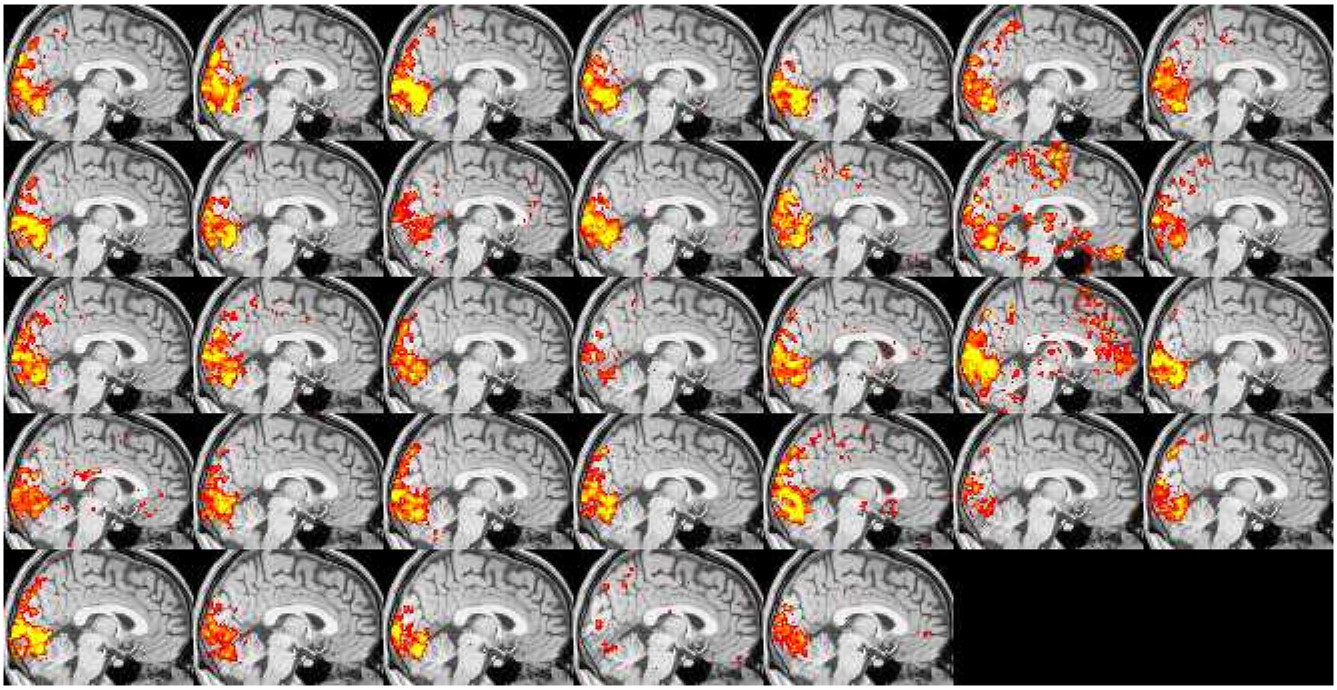

Figure 7: Visual paradigm; analysis F single-session thresholded maximum intensity projections, thresholded with the t threshold reduced from the "p<0.05 GRF-corrected" level by 33%. Note that as well as the obvious increase in reported activation, "apparent variability" is significantly decreased.

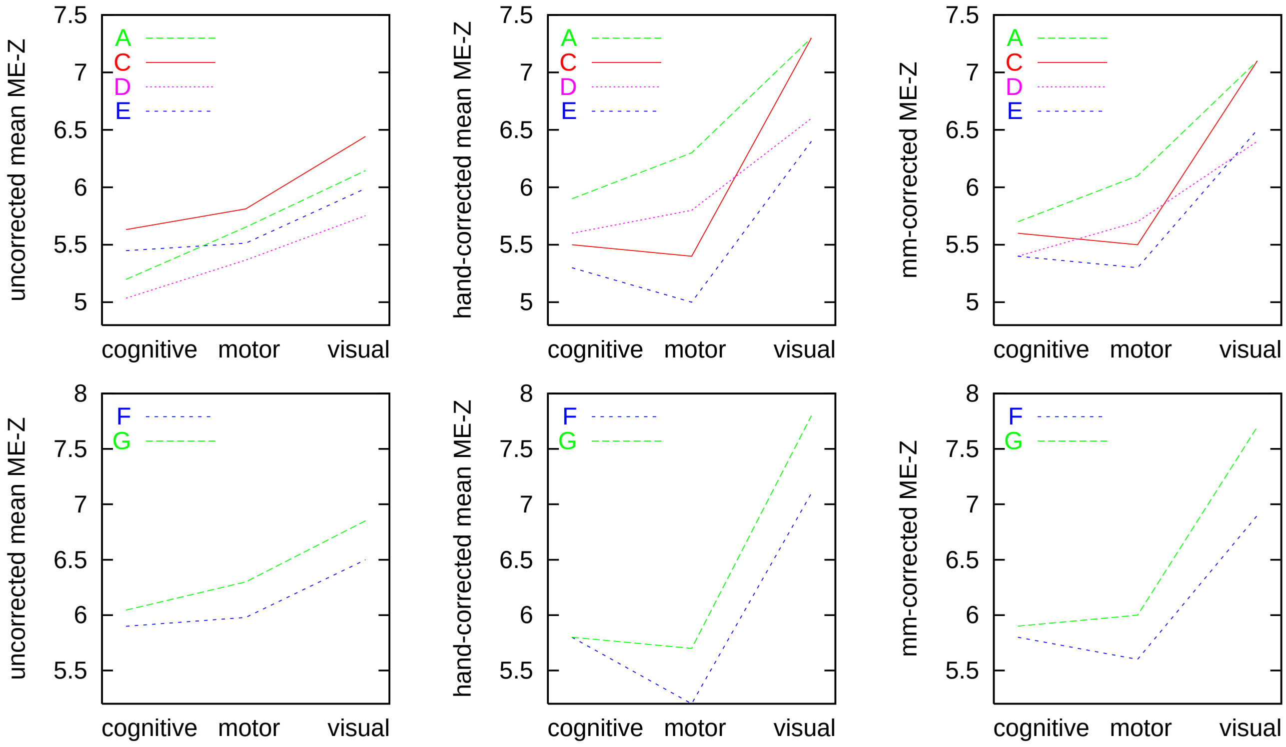

Figure 8: Mean mixed-effects Z values, uncorrected and with both z-shift correction methods.

the original analyses in [18], including turning on intensity normalisation in both cases.

The results show that both time-series statistics and spatial components (primarily head motion correction and registration to standard space) contribute to (i.e., add to) apparent session variability. Overall, with respect both to spatial alignment processing and time series statistics, FSL induced less error than SPM, i.e., was more efficient with respect to higher-level activation estimation.

The experiments used for this paper used a block-design, and as such are not expected to show up the increased estimation efficiency of prewhitening over precolouring [22]. In, [5], a similar study to that presented here, first-level statistics were obtained using SPM99 and FSL (i.e., only time-series statistics were compared, not different alignment methods). The data was primarily event-related, and, as in this paper, simple second-level mixed-effects analysis was used to compare efficiency of the different methods. The results showed that prewhitening was not just more efficient at first-level, but also gave rise to increased efficiency in the second-level analysis.

4.7 Intensity Normalisation

FEAT offers the option of intensity normalisation (of all volumes in each time series in order to give constant mean volume intensity over time); however, this option is turned off by default, as it is considered that this is an over-simplistic approach to a complicated problem (see, for example, [7]).

We investigated the effect (on inter-session variance) of turning intensity normalisation on. It was found that this pre-processing step does reduce the overall fixed and random effects variance (on average by about 10%), and therefore slightly increases the fixed and random effects Z values (again, giving, on average, approximately a 10%

increase in the number of supra-threshold voxels).

4.8 1- or 2-Step Registration

FEAT does not transform the FMRI data directly into standard space but carries all statistics out in the original (low resolution) space and then transforms the final statistics images into standard space. The transformation from original space into standard space is normally carried out (automatically) in a two-step process; first an example functional image (the one which was used as the reference in the motion correction) is registered to the subject's structural image (normally a T1-weighted image which has been brain-extracted using BET [20]) and then the structural image is registered to a standard space template (normally the MNI152). The two resulting transformations are concatenated resulting in a single transform which takes the low resolution statistic images into standard space. This is the default FEAT registration procedure, and is what was used for the analyses presented above.

We investigated whether, for this data, FEAT's two-step process (using FLIRT) is an improvement over registering the example functional image directly into standard space (using FLIRT). The two-step registration resulted in a slight decrease in cross-session fixed and random effects overall variance (by approximately 3%). The number of activated voxels in general stayed the same, but the peak Z statistic improved (again by approximately 3%) when two-step registration was used, and the activation appeared qualitatively to contain more structural detail (i.e., was less blurred). The conclusion therefore is that, even for this within-subject across-session analysis, the two-step registration approach was of value in the FEAT analyses.

5 Conclusions

Intersession variability is an important consideration in power calculations for the design of FMRI experiments. It is also a critical issue for the interpretation of studies that allow for only single observations, e.g., in many clinical applications of FMRI. Here we have provided quantitative data confirming that intersession variability in FMRI is not large relative to within-session variability. We also emphasise that inter-session variability should not be judged by apparent variability in thresholded activation maps.

There are several mechanisms by which intersession variability can be minimised. While considerable attention has been paid in the past to hardware and experimental design factors, we have shown here that additional benefits can come with optimisation of analysis methodology, as analysis methods add extra variance to the true intersession variance, causing apparent increase in intersession variance. It was found that with respect both to spatial alignment processing and time series statistics, FSL1.3 induced less error than SPM99b, i.e., was more efficient with respect to higher-level activation estimation.

6 Acknowledgements

We are grateful for support from the Medical Research Council (UK), the Engineering and Physical Sciences Research Council (UK), the EPSRC Medical Images and Signals Collaboration and GSK. We are very grateful to Chris Freemantle for the retrieval and transfer of the data from the Functional Imaging Lab, London.

References

[1] C.F. Beckmann, M. Jenkinson, and S.M. Smith. General multi-level linear modelling for group analysis in FMRI. *NeuroImage*, 20:1052–1063, 2003.

[2] C.F. Beckmann, M.W. Woolrich, and S.M. Smith. Gaussian / Gamma mixture modelling of ICA/GLM spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*, 2003.

[3] T. Behrens, M.W. Woolrich, and S.M. Smith. Multi-subject null hypothesis testing using a fully Bayesian framework: Theory. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*, 2003.

[4] R. Beisteiner, C. Windischberger, R. Lanzenberger, V. Edward, R. Cunnington, M. Erdler, A. Gartus, B. Streibl, E. Moser, and L. Deecke. Finger somatotopy in human motor cortex. *NeuroImage*, 13:1016–1026, 2001.

[5] M. Bianciardi, A. Cerasa, and G. Hagberg. How experimental design and first-level filtering influence efficiency in second-level analysis of event-related fMRI data. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*, 2003.

[6] M.W.L. Chee, H.L. Lee, C.S. Soon, C. Westphal, and V. Venkatraman. Reproducibility of the word frequency effect: Comparison of signal change and voxel counting. *NeuroImage*, 18:468–482, 2003.

[7] M. De Luca, C.F. Beckmann, T. Behrens, S. Clare, P.M. Matthews, N. De Stefano, M. Woolrich, and S.M. Smith. Low frequency signals in FMRI - ”resting state networks” and the ”intensity normalisation problem”. In *Proc. Int. Soc. of Magnetic Resonance in Medicine*, 2002.

[8] A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly, and T.M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Proc. IEEE-nuclear science symposium and medical imaging conference*, pages 1813–1817, 1993.

[9] G. Fernandez, K. Specht, S. Weis, I. Tendolkar, M. Reuber, J. Fell, P. Klaver, J. Ruhlmann, J. Reul, and C.E. Elger. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, 60(6):969–75, 2003.

[10] K.J. Friston, O. Josephs, E. Zarahn, A.P. Holmes, S. Rouquette, and J-B. Poline. To smooth or not to smooth? *NeuroImage*, 12:196–208, 2000.

[11] K.J. Friston, K.J. Worsley, R.S.J. Frackowiak, J.C. Mazziotta, and A.C. Evans. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220, 1994.

[12] A.P Holmes and K.J. Friston. Generalisability, random effects & population inference. In *Fourth Int. Conf. on Functional Mapping of the Human Brain, NeuroImage*, volume 7, page S754, 1998.

[13] M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[14] M. Jenkinson and S.M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, June 2001.

[15] H. Johansen-Berg, H. Dawes, C. Guy, S.M. Smith, D.T. Wade, and P.M. Matthews. Correlation between motor improvements and altered fMRI activity after rehabilitative therapy. *Brain*, 125(12):2731–42, 2002.

[16] D.G. Leibovici and S. Smith. Comparing groups of subjects in fMRI studies: a review of the GLM approach. Technical Report TR00DL1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK, 2000. Available at `www.fmrib.ox.ac.uk/analysis/techrep` for downloading.

[17] H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

[18] D.J. McGonigle, A.M. Howseman, B.S. Athwal, K.J. Friston, R.S.J. Frackowiak, and A.P. Holmes. Variability in fMRI: An examination of intersession differences. *NeuroImage*, 11:708–734, 2000.

[19] M.E. Shaw, S.C. Strother, M. Gavrilescu, K. Podzebenko, A. Waites, J. Watson, J. Anderson, G Jackson, and G. Egan. Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. *NeuroImage*, 19:988–1001, 2003.

[20] S.M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, November 2002.

[21] L.G. Ungerleider, J. Doyon, and A. Karni. Imaging brain plasticity during motor skill learning. *Neurobiol. Learn. Mem.*, 78(3):553–64, 2002.

[22] M.W. Woolrich, B.D. Ripley, J.M. Brady, and S.M. Smith. Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386, 2001.

[23] K.J. Worsley and K.J. Friston. Analysis of fMRI time series revisited - again. *NeuroImage*, 2:173–181, 1995.