The effects of single-trial averaging upon the spatial extent of fMRI activation

Scott A. Huettel[1,CA] and Gregory McCarthy[1,2]

[1]Brain Imaging and Analysis Center, Duke University Medical Center, Box 3918, Durham, NC 27710; [2]Department of Veterans Affairs Medical Center Durham, NC, USA

[CA]Corresponding Author

We examined effects of trial averaging upon spatial extent, spatial topography, and temporal properties of fMRI activation. Two subjects participated in an event-related visual stimulation design. There was an exponential relation between number of trials and spatial extent, such that additional trials identified, on average, a constant proportion of the remaining voxels. At values typical of fMRI experimentation (e.g. 50 trials) only about 50% of eventually active voxels were significant; asymp- totic values were approached by 150 trials. The variability of the estimated hemodynamic response decreased with signal averaging, becoming stable across samples of ≥ 25 trials. Therefore, group or condition differences may result from differences in voxelwise noise exacerbated by averaging small numbers of trials. *NeuroReport* 12:1–6 © 2001 Lippincott Williams & Wilkins.

Key words: fMRI; Functional magnetic resonance imaging; Hemodynamic response; Noise; Signal averaging

INTRODUCTION

The spatial extent of activation is a critical dependent variable in many fMRI studies. For example, in presurgical planning, the spatial extent of a language or motor activa- tion relative to the location of a tumor helps to determine the margins of a resection [1]. In studies of motor training, the differences in the spatial extent of activation between trained and untrained states have been interpreted in terms of cortical plasticity [2]. In between-group studies, such as those between hyperactive and normal subjects [3], or between the elderly and young [4], differences in spatial extent of activation may be interpreted in terms of hypoac- tive or dysfunctional cortex.

Recent studies have found that the spatial extent of fMRI activation in healthy older adults is approximately half that of younger subjects, for both visual [5] and visuomotor [6] tasks. Importantly, this difference is not associated with reduced hemodynamic response (HDR) amplitude in older subjects, because young and elderly adults have similar HDR amplitudes [5–7], and have similar distributions of HDR amplitudes across voxels [5]. Furthermore, greater head motion in the elderly does not cause this difference [5,7], although head motion differences can affect spatial extent of activation [8]. Instead, spatial extent differences are associated with higher voxelwise noise levels in elderly adults [5–7], perhaps due to changes in cardiac or respira- tory effects upon the fMRI signal [7].

Signal averaging improves the signal-noise ratio (SNR) as an approximate function of the square root of the number of trials averaged. So, if young and elderly adults participate in similar experimental designs, a difference in spatial extent of activation will be measured. For example, HDRs of similar amplitude may exceed a given statistical threshold for younger subjects but not for elderly subjects. Averaging more trials for the elderly, which would have increased their effective SNRs and thus increased the spatial extent of activation, could have ameliorated this difference. Simulation results have supported this conclu- sion, finding that empirical differences in spatial extent of activation were consistent with predicted results due solely to SNR differences between groups, and that averaging larger numbers of trials reduced spatial extent differences [5].

That one can improve the SNR of a weak signal within noise by signal averaging and thus improve its detection is hardly a novel insight. What is surprising, however, is the degree of variation in spatial extent that occurs when working within the typical ranges of trial numbers and SNR of event-related fMRI studies [5]. The number of trials averaged differs greatly in the published literature, and it is uncommon to report voxelwise SNR values. However, in the absence of precise knowledge of SNR differences be- tween groups, between brain regions in the same subject, or between different categories of stimuli, the interpreta- tion of differences in the spatial extent of fMRI activation is a risky enterprise.

As our assertion that the number of trials averaged influences the spatial extent of the fMRI activation in typical studies was based upon a *post-hoc* simulation [5], we conducted an empirical study in which two subjects were extensively tested in a visual event-related fMRI paradigm. Here we report the investigation of three issues.

First, we investigated the relation between the number of trials averaged and the number of active voxels measured. Our goal was to examine this relation in sufficient detail so that researchers can make informed decisions about how many trials to include in experimental designs. Second, we investigated whether the topography of activation, and not just its spatial extent, changes with trial averaging. Finally, we investigated the variability in the amplitude and form of the HDR across many samples of averages randomly drawn from our empirical population of single trial HDRs.

MATERIALS AND METHODS

Subjects: Two right-handed male subjects (S1 and S2) were tested (ages S1: 47 years; S2: 34 years). Both subjects had corrected to normal vision and were experienced with fMRI studies. The experimental protocol was approved by the Institutional Review Board of Duke University and written informed consent was obtained.

Stimuli and experimental design: Stimuli were black and white radial checkerboards that subtended $20 \times 15°$ of visual angle. The stimuli were projected on a screen directly behind the subject's head within the scanner bore. Subjects viewed the stimuli with mirrored glasses. Each checkerboard was presented singly for 500 ms. A fixation cross was visible during the interstimulus interval, which varied randomly between 14 and 18 s. The number of single trials acquired was 192 for S1 and 154 for S2. The trials were presented within runs lasting ~6 min, with short breaks (< 30 s) between runs.

Imaging parameters: Scans were acquired on a GE Signa NVi 1.5 T scanner equipped with 41 mT/m gradient coils. Slice selection followed the acquisition of sagittal scout images. In S1, 10 contiguous 5 mm slices were acquired parallel to the line connecting the anterior and posterior commissures (axial imaging plane). In S2, 12 contiguous 5 mm slices were acquired perpendicular to the line connecting the anterior and posterior commissures (coronal imaging plane). These different imaging planes were chosen to investigate topography changes in visual cortex activation. High-resolution T1-weighted spin-echo images were collected at each slice location (in-plane resolution 0.94 mm²). Functional gradient echo-planar images were acquired at a TR of 1 s (TO 40 ms, flip angle 81°, FOR 24 cm, matrix: 64², in-plane resolution 3.75 mm²).

Data analysis: All analyses were conducted using custom MATLAB scripts written by the authors. Single-trial epochs, defined from 5 time points preceding through 13 time points following checkerboard onset, were excised from the continuous time series acquisition. These single trial populations; were then randomly sampled and a series of signal averages varying in the number of single trials combined were computed for each subject. Twenty signal averages were computed for each possible number of trials (192 for S1, 154 for S2) so that HDR variability could be measured. Note that since 192 trials constituted the entire population for S1, each of the 20 samples of 192 trials was identical. Therefore, we were unable to estimate HDR variability when the numbers of trials in the sample approached the size of the population.

For each signal average, the average time epoch was correlated with an empirical HDR function [9]. This function had a peak latency of 5 s; however, in different analyses below the HDR was lagged so that its peak latency varied from 4 to 6 s in 1 s increments. The threshold for significance for the correlation was set at $t > 3.6$ ($p < 0.001$, uncorrected). The number of voxels exceeding that criterion was determined for each of the 20 samples at each number of trials for each subject. To simplify the analysis and avoid potential biases due to the interleaved acquisition of slices within each TR, these counts were performed within an anatomically defined region of interest (ROI) within a single slice for each subject. For both subjects, the ROI included peri-calcarine and ventral extrastriate cortex.

From these analyses, the SNR was determined for each voxel in the following manner. SNR was defined as the peak amplitude of the averaged HDR response divided by the standard deviation of the signal variation in that voxel over the entire continuous time series (equivalent to the z-score of the peak HDR from the temporal variation at each voxel). The contribution of the HDR itself to signal variation was removed by subtracting the mean HDR response in each voxel from all single trial epochs.

Finally, we examined the effects of trial averaging upon the stability of the hemodynamic response. An active set of voxels was defined as those within each ROI that were significant following averaging of all trials for that subject. The mean HDR in that ROI was calculated for all trial numbers averaged. The standard deviation was computed for the amplitudes at each time point across the 20 estimates for each of the trial numbers averaged.

RESULTS

Number of active voxels: The number of activated voxels exceeding criterion is plotted against the number of trials in each average in Fig. 1. As predicted, the spatial extent of activation increased with increasing number of trials. The growth in spatial extent followed a pattern similar to that found in our prior simulation [5]. Adding additional trials when only few trials were averaged greatly changed the spatial extent. Both curves approached (but did not reach) an asymptote for large numbers of trials, suggesting that most of the active voxels had been identified.

We found that an exponential function fit the form of these curves. This function, shown in eqn 1, indicated that the proportion of missed active voxels decreases in an exponential fashion as the number of averaged trials increases. The number of total voxels, V_{max}, was estimated independently in each subject by minimizing the sum of squared deviations, since the results suggested that we had not reached asymptote in either subject even after averaging more than 150 trials (V_{max} values in Fig. 1 inset).

$$V_N = V_{max}[1 - e^{(-0.016 \times N)}] \tag{1}$$

Equation 1 accounted for 99% of the variance in the data for both subjects. The deviations that were observed between the data and the prediction of eqn 1 were in opposite directions for S1 and S2. For small numbers of trials, S1 had slightly more voxels active than predicted, while S2 had slightly fewer voxels active than predicted.
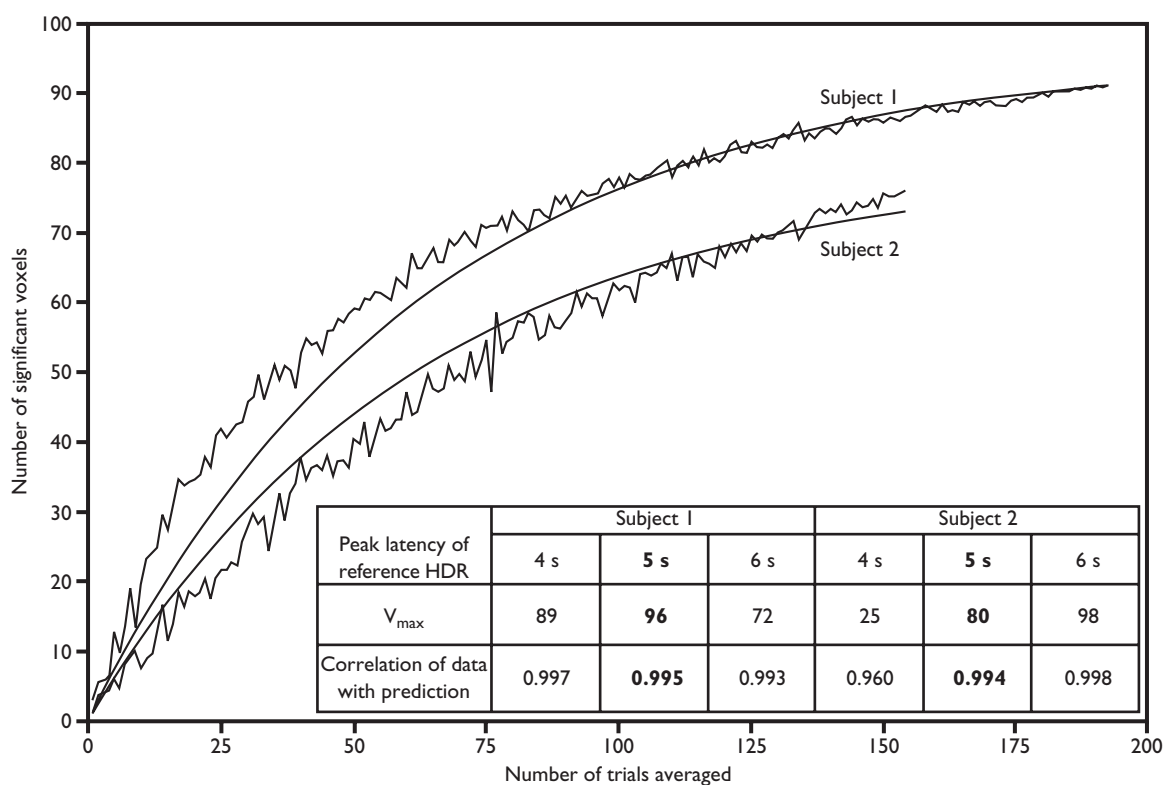
Fig. 1. Dependence of the spatial extent of fMRI activation upon the number of trials averaged. Two subjects were tested in a passive viewing task, in which checkerboard stimuli were presented at fixation (see text for details). Shown, in jagged lines, are the numbers of significant voxels determined by correlation to a reference hemodynamic response with peak at 5 s, recorded from samples of 1–192 trials (S1) and of 1–154 trials (S2). For both subjects, the empirical data closely fit an exponential function (eqn 1 in text). The maximum number of voxels identified and the fit of the empirical data to eqn 1 are indicated in the inset table, for hemodynamic responses with peaks at 4, 5, and 6 s.

Examination of the results across all HDR lags tested suggested that this discrepancy was due to slight differences in mean HDR latency between the subjects. The correlation between eqn 1 and the empirical data increased to 0.997 for S1 when a reference HDR with a 4 s peak was used in the analysis. Similarly, the correlation increased to 0.998 for S2 when a HDR peak of 6 s was used. Simulation testing using methods similar to those reported by Huettel and colleagues [5] indicated that eqn 1 is consistent with signal amplitude having a gamma distribution across voxels and noise amplitude having a normal or supranormal distribution across time.

Topography of activation: We investigated whether there were systematic changes in the topography of activation with increasing numbers of averaged trials, which would suggest that the voxelwise SNR was not randomly distributed throughout the ROI. This might reflect systematic changes in the amplitude and latency of the HDR across or between functional regions, or systematic differences in noise that might reflect partial voluming of different tissue types at gray/white matter boundaries.

Figure 2 shows activation maps for the two subjects for different numbers of trials averaged. To simplify the figure, only averages computed for 4, 16, 3 6, 64, 100 and 144 trials are shown, corresponding to proportional improvements in SNR of 2, 4, 6, 8, 10 and 12. When few trials were averaged

(e.g. 16), the active voxels were clustered near the center of the eventual activation region. As additional trials were averaged, the activation changed from medial/posterior fusiform to include more anterior and lateral areas in S1, and to include inferior and lateral areas (including the lingual gyrus) in S2.

Variability of the hemodynamic response: The ability to detect differences in the amplitude of the HDRs evoked by different stimuli at the same voxel is dependent upon the variability of the HDR amplitude across trials. Figure 3 presents HDR variability for averages across different numbers of trials.

It is interesting to note that many single trials evoked recognizable HDRs, a result first reported by Blamire and colleagues [10]. The HDR is clearly evident in the averages of 4–16 trials; however, there is considerable variability in amplitude and form across samples. By 25 or more trials the form of the HDR varied little from sample to sample. This is shown quantitatively on the standard deviation plot in Fig. 3, lower right, which presents s.d. across samples at HDR peak (5 s after stimulus onset). Similar curves were observed at every time point in the epoch. To ensure that our random sampling technique, which allowed the same trial to be a part of multiple samples, did not introduce bias, we conducted two analyses of randomly generated data with the same mean and variance as our empirical
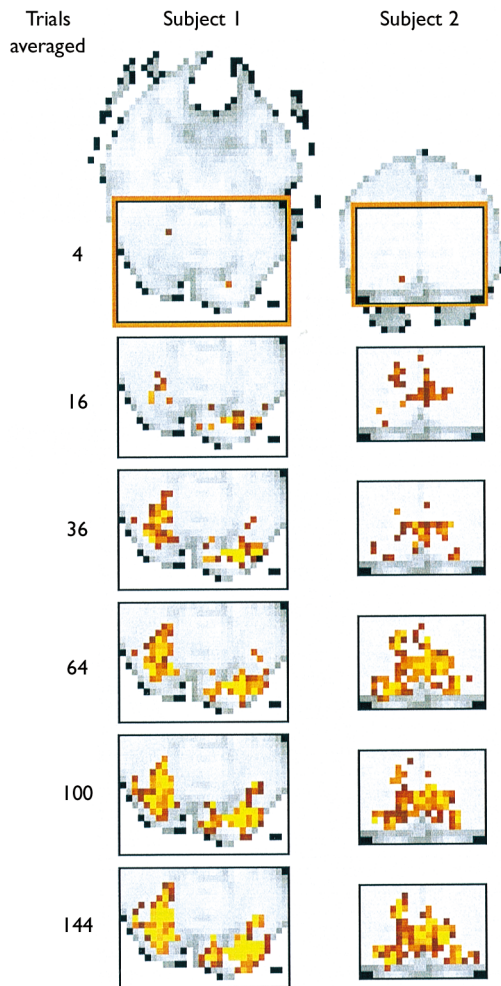
Fig. 2. Changes in the spatial topography of activation with increasing numbers of trials. Shown are functional activation plots, for both subjects tested, derived from samples of 4, 16, 36, 64, 100, and 144 trials. These values were chosen to represent improvements in SNR by factors of 2, 4, 6, 8, 10, and 12, respectively. As the number of trials increases, not only does the total number of active voxels increase, but also the pattern of activation changes to include voxels more removed from the center of activation. For S1, the extent of activation spreads anteriorly and laterally, and for S2, the extent of activation spreads inferiorly and laterally.

data. The first analysis used 192 trials, from which 20 samples of N trials were extracted from the set for calculation of the standard deviation. The second set consisted of completely independent trials for all samples. Notably, the effects of overlapping samples were only evident at sample sizes of 49 and greater, indicating that our above conclusion was not biased.

DISCUSSION

These empirical results confirmed our prior simulation [5], indicating that the spatial extent of fMRI activation depends strongly upon the number of trials averaged. The appropriate number of trials for an fMRI experiment depends on its goals, whether activation detection or HDR estimation [11]. Our results suggest that, if an experiment is intended to detect whether there are active voxels within

an area, then a relatively small number of trials (< 20) may be sufficient. For example, detectible changes in fMRI activation have been reported with single trials [12]. But, if the experiment is intended to determine the spatial extent of activation by detecting all or nearly all active voxels, then many more trials (> 100) should be averaged, far more than typical for fMRI studies. Conversely, if the goal is to estimate the HDR from a identified region, then relatively fewer trials (25–36) are required.

Several issues remain as targets for future studies. First, given the considerable interest in high-field fMRI research, it will be critical to examine how spatial extent changes with number of trials as a function of field strength. We expect that the general relation identified here, as represented by an exponential function, will hold across field strengths, but that the specific parameters in that relation may differ. Higher SNR in individual voxels at 4.0 T, for example, would reduce the number of trials needed to detect any voxel, decreasing the exponent in eqn 1.

Second, the degree to which fMRI noise exhibits spatial autocorrelation requires examination. If the voxel-wise noise was Gaussian, then spatial smoothing might improve SNR and reduce the number of trials required for averaging. However, if some components of the noise were correlated among adjacent voxels, then spatial smoothing would be less useful. No spatial smoothing was performed in this study.

The distribution of noise across the brain is not uniform; indeed, an image that reflects the standard deviation of activity within each voxel's unstimulated time series shows considerable spatial structure. In general, gray matter has a higher noise level than white matter. This structure suggests that the composition of any given voxel will have significant effects upon underlying noise [13]. Partial volume effects, as when a voxel contains both gray and white matter, will influence both task-related signal and voxel-wise noise. In voxels that include, or are near, high noise sources like blood vessels, averaging few trials may be insufficient. Conversely, voxels in more stable regions may have less stringent requirements for signal averaging. We suggest that power analyses, which are typically conducted on subject number or, less frequently, trial number, should be considered for individual voxels.

Finally, a central issue in fMRI analysis lies in understanding the sources of voxelwise noise. Within a single subject, noise levels differ according to voxel location, with voxels nearer to tissue boundaries being more susceptible to changes in signal intensity due to head motion. Likewise, voxel contents, whether gray matter, white matter, ventricle, or some combination, influence noise level. Pulse sequences that reduce the signal contributions from certain brain structures, such as diffusion-weighting techniques that attenuate signal from large blood vessels, will modify the effects of signal averaging in affected voxels. Other possible sources of noise, such as effects of heart or respiratory rate that may differ across the brain, are less well understood. Across-subjects comparisons introduce many other sources of noise, especially when the subject samples compared are drawn from different populations (e.g. young and elderly). Physiological changes in hemodynamic responsiveness may accompany aging or disease states, influencing noise levels. As more fMRI studies
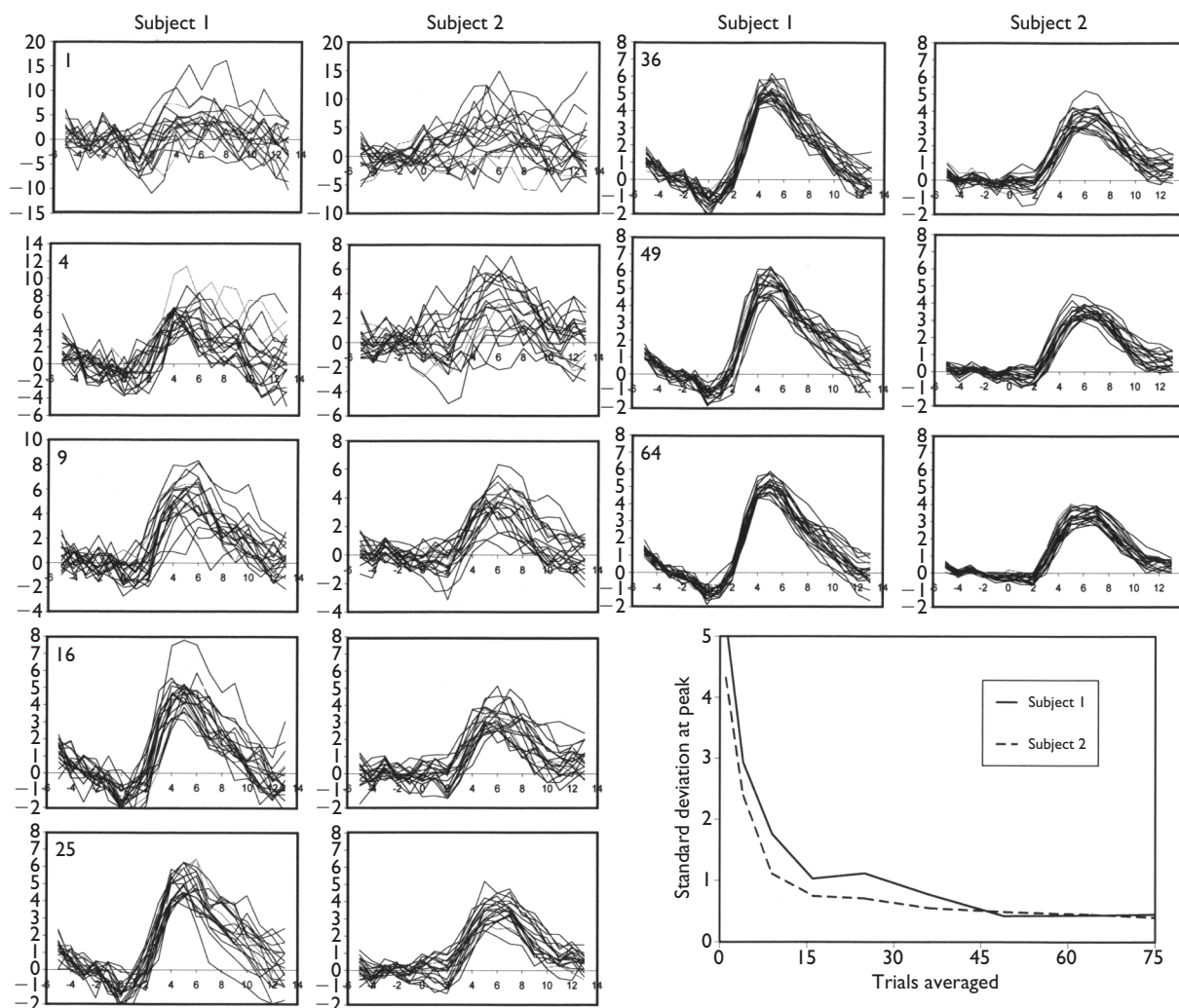
Fig. 3. Dependence of the form of the hemodynamic response upon the number of trials averaged. Each graph shows a set of 20 HDR functions derived by averaging activation in a region of interest over N trials. Note that the y-axis scales on the graphs for averages of 1, 4, and 9 trials are larger than the others due to their greater variability. At top left are single-trial HDRs, which show considerable variability. As more trials are averaged, stability of the HDR greatly increases across samples. The graph at lower right indicates the effects of trial averaging upon the standard deviation of the HDR peak (at 5 s). Notably, increasing the number of trials to around 25 has large effects on HDR stability, whereas further increases provide less improvement.

investigate group differences, an understanding of what factors contribute to fMRI noise will be necessary for accurate reporting of differences in spatial extent of activation.

CONCLUSION

Our results indicate that spatial extent of fMRI activation depends upon the SNR of the HDR, which in turn varies across space within subjects. This finding, though theoretically consistent with the basis for signal averaging, has practical implications for any study that wishes to interpret between-group differences in fMRI spatial extent as indicating differences in functional states of cortex. Any systematic difference between groups, such as young and elderly adults, in noise properties of the HDR will lead to differences in spatial extent of activation at trial numbers

typical for fMRI experiments (<100). This finding has implications for within-subject designs as well. Noise differences induced by experimental manipulations, such as between drug and placebo conditions, may also cause differences in spatial extent of activation. Likewise, SNR improvements associated with signal averaging may change the topography of fMRI activation across conditions. Finally, the form of the fMRI HDR is stable across repeated samples following averaging of $\geqslant 25$ trials, suggesting that HDR form can be reliably estimated within typical experimental designs.

REFERENCES
1. Schlosser MJ, Luby M, Spencer DD *et al*. *J Neurosurg* **91**, 626–635 (1999).
2. Kami A, Meyer G, Jezzard P *et al*. *Nature* **377**, 155–158 (1995).
3. Bush G, Frazier JA, Rauch SL *et al*. *Biol Psychiatry* **45**, 1542–1552 (1999).

4. Ross MH, Yurgelun-Todd DA, Renshaw PF *et al. Neurology* **48**, 173–176 (1997).
5. Huettel SA and McCarthy G. *NeuroImage* **13**, 161–175 (2001).
6. Buckner RL, Snyder AZ, Sanders AL *et al. J Cogn Neurosci* **12**, 24–34 (2000).
7. D'Esposito M, Zarahn E, Aguirre GK *et al. NeuroImage* **10**, 6–14 (1999).
8. Callicott JH, Ramsey NF, Tallent K *et al. Neuropsychopharmacology* **18**, 186–196 (1998)
9. Huettel SA and McCarthy G. *NeuroImage* **11**, 547–553 (2000).
10. Blamire AM, Ogawa S, Ugurbil K *et al. Proc Natl Acad Sci USA* **89**, 11069–11073 (1992).
11. Liu TT, Frank LR, Wong EC *et al. NeuroImage* **13**, 759–773 (2001).
12. Richter W, Somorjai R, Summers R *et al. J Cogn Neurosci* **12**, 310–320 (2000).
13. Parrish TB, Gitelman DR, LaBar KS *et al. Mag Reson Med* **44**, 925–932 (2000).