Early Stopping in Experimentation with Real-time Functional Magnetic Resonance Imaging Using a Modified Sequential Probability Ratio Test

Sarah J. A. Carr[1,2], Weicong Chen[3], Jeremy Fondran[4], Harry Friel[5], Javier Sanchez-Gonzalez[6], Jing Zhang[4] and Curtis Tatsuoka[4,2,*]

1. Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK
2. Department of Neurology, Case Western Reserve University, Cleveland, OH, USA
3. Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, USA
4. Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA
5. Philips Healthcare, Highland Heights, OH, USA
6. Philips Healthcare, Madrid, Spain

*Corresponding Author:
Curtis Tatsuoka
10900 Euclid Avenue
Case Western Reserve University
Cleveland, OH, USA 44106
Email: cmt66@case.edu

KEY WORDS: real-time fMRI, adaptive fMRI, dynamic experimentation, SPRT, early stopping

Abstract

Introduction: Functional magnetic resonance imaging (fMRI) often involves long scanning durations to ensure the associated brain activity can be detected. However, excessive experimentation can lead to many undesirable effects, such as from learning and/or fatigue effects, discomfort for the subject, excessive motion artifacts and loss of sustained attention on task. Overly long experimentation can thus have a detrimental effect on signal quality and accurate voxel activation detection. Here, we propose dynamic experimentation with real-time fMRI using a novel statistically-driven approach that invokes early stopping when sufficient statistical evidence for assessing the task-related activation is observed.

Methods: Voxel-level sequential probability ratio test (SPRT) statistics based on general linear models (GLMs) were implemented on fMRI scans of a mathematical 1-back task from 12 healthy teenage subjects and 11 teenage subjects born extremely preterm (EPT). This approach is based on likelihood ratios and allows for systematic early stopping based on statistical error thresholds. We adopt a two-stage estimation approach that allows for accurate estimates of error variance. Early stopping performance is reported for different first stage lengths, and activation results are compared with full durations. Finally, group comparisons are conducted with both early stopped

and full duration scan data. Numerical parallelization was employed to facilitate completion of computations involving a new scan within every repetition time (TR).

Results: Use of SPRT demonstrates the feasibility and efficiency gains of automated early stopping, with comparable activation detection as with full protocols. Dynamic stopping of stimulus administration was achieved in around half of subjects, with typical time savings between 33 - 66% (4 – 8 minutes on a 12 minute scan). A group analysis produced similar patterns of activity for control subjects between early stopping and full duration scans. The EPT group, individually, demonstrated more variability in location and extent of the activations compared to the normal term control group. This was apparent in the EPT group results, reflected by fewer and smaller clusters.

Conclusion: A systematic statistical approach for early stopping with real-time fMRI experimentation has been implemented. This dynamic approach has promise for reducing subject burden and fatigue effects.

1.0 Introduction

Analysis of task-based fMRI scans is typically performed with fixed, predetermined experimental designs. As a result, subjects must often endure stimulus protocols that are overly long in order to ensure the neural activity can be statistically discerned in the noisy data. However, this can lead to fatigue, learning effects and excessive motion, such as from agitation, as well as being costlier to administer due to longer scan times and potentially less reliable measurement. Also, the experimenter does not know if the neural activity is detectable until long after the scanning session is over. Real-time functional MRI (RT-fMRI) provides an opportunity to ameliorate these issues. RT-fMRI has been successfully applied in the field of neurofeedback and biofeedback from neural responses, where subjects may be trained to alter their brain activity based on real-time information provided from the fMRI scans. This has been reported in ADHD (1), healthy subjects with no psychiatric or neurological disorders (2, 3), Alzheimer's disease (4) and Parkinson's disease (5, 6). Its uses have also been described in psychoradiology to aid diagnosis and treatment planning in psychiatric disorders (7). A largely unexplored application of RT-fMRI is to dynamically and statistically determine when a stimulus has been sufficiently presented in terms of replication of blocks to terminate early. The magnitude of effort and variability in neural activity while completing a task will vary from person to person. Trial administration within a block design can be stopped early if sequentially updated statistical inference on activation can be determined with sufficient accuracy based on the observed BOLD (blood oxygen level dependent) signal response up to that point. This application will be explored in detail.

The benefits of adaptive RT-fMRI include: 1) <u>Shorter scan times for fMRI testing</u>: Shorter scan times cannot only save in technology and personnel costs, but fatigue and learning effects can be avoided, improving signal quality. Scanning becomes less burdensome on the subject as well, which is an especially important consideration for children or elderly subjects. 2) <u>Real-time quality control</u>: greater consistency in activation classification error can be obtained, through statistical error-based benchmarks for stopping rules and real-time feedback on classification

performance and adjustment of stimulus durations. 3) Richer information: Paradigms can become more complex and sophisticated. With greater time efficiency and flexibility, more variations of a stimulus, such as reflected by a broader range of difficulty levels, can be administered in the same amount of time. 4) Wide applicability: Dynamic adjustment of stimuli based on BOLD response in real time can be generally applied across a range of focus areas that investigate localization of brain activity, including cognition and motor functioning.

Since the advent of RT-fMRI in the mid 1990's (8), a handful of mainstream software packages have been developed for use by the fMRI community. These include Turbo BrainVoyager (9), AFNI's real-time plugin (8) and FSL-based FRIEND (10). There have been few previous studies that have used adaptive RT-fMRI. In one example, it has been used to determine 'good' and 'bad' brain states to optimize learning (11). The presentation of novel scenes was prompted by the detection of 'good' brain states; the 'good' template was determined based on a prior standard acquisition test scan. They used real-time general linear model (GLM) methods described in (12) to estimate the BOLD signal magnitude at each time point (each scan) and compared it to a value within a region of interest from the earlier test scan. Another adaptive RT-fMRI study has used a person's brain state to judge their attention to a task (13). When their attention appeared to wander, the difficulty of the task was increased bringing their attention back. The authors used multivariate pattern analysis to determine task-relevant and task-irrelevant activity. Lorenz et al (2016) used FSL to pre-process the scans in real-time before applying a GLM-based analysis. Their study involved eliciting activity in particular brain regions by presenting stimuli chosen based on the response to the previous stimulus. The aim was not to investigate brain activity related to a particular task but simply activate a brain region (14). Another example of adaptive RT-fMRI used a Bayesian optimization algorithm to estimate when brain activity was mapped to a particular network (15). The Bayesian optimization was trained on 4 difficulty levels of a task prior to switching to choosing the optimal difficulty levels to elicit the desired activity, where there were 12 other levels to choose from.

Here, we extend the use of a statistically-based dynamic approach to RT-fMRI experimentation described in (16). This approach involves the sequential updating of voxel-level likelihood ratio tests, known as sequential probability ratio tests (SPRTs) and assessing after each scan whether there is sufficient statistical evidence to determine whether or not an associated parameter value indicates task activation. Such results, considered in aggregate across a collection of voxels, can be used as a basis for early stopping of experimentation. More generally, this approach can apply to competing hypotheses related to linear contrasts of task parameters. Our extension involves the actual implementation of this system on a Philips MRI scanner, in a real-time workflow, and sequential analysis of subjects across a study. This allows us to assess the variability in experimentation lengths across individuals to reach a rigorous threshold for decisive classification of activation. This work supports the premise that adaptive, individualized experimentation is feasible and can lead to savings in experimental durations.

Most off-line, post-hoc analyses of fMRI data use the general linear model to test statistical associations of voxel activation magnitude to task administration (17-19). This approach involves the voxel-level estimation of task-related regression parameters that indicate magnitude of

association between an expected hemodynamic response signal from a task and the observed BOLD signal. In our previous work, we adapted this general method for real-time fMRI by sequentially updating GLM regression parameter estimates as soon as the brain volumes were collected. At the individual voxel level, we can then assess hypothesis tests related to activation that are based on these estimates. In aggregate, the voxel level analyses inform decisions on early stopping and the tailoring of fMRI experimentation. See (16) for more details.

In comparison to (16), we previously assumed independence when estimating the covariance model. Here, we extend these methods by considering the "sandwich" estimator to recognize serial covariance (20, 21). We also adopt a two-stage estimation approach that allows for the hypothesis test parameters to be formulated in terms of z-score scale at the voxel level. Moreover, we now present a novel workflow to apply and implement these methods on a Philips scanner, with a dynamic feedback system that allows for real-time dynamic adjustment of the experimentation with subjects. This was facilitated with adoption of numerical parallelization techniques.

Another novel aspect of this work is the application of adaptive RT-fMRI in a sample group of 12 healthy adolescent subjects and 11 adolescents born extremely preterm (EPT). Early stopping was implemented using sequential probability ratio test (SPRT) statistics and our server was a custom-built Linux computer located in a nearby building. The fMRI stimulus was a mathematical version of the well-known 1-back task. We show a novel workflow for the real-time processing of fMRI scans; a workflow that allows adaption of the stimulus difficulty level; processing of RT-fMRI was completed within 3 seconds before the next scan arrived; average time savings of 33 - 60% based on early stopping when 80% of voxels were classified. This equated to 4 – 8 minutes savings with a 12-minute scan. The impact on activation analysis from selecting the early stopping criteria of 80% of voxels is assessed, as described in detail below. Further, we review motion events over the scan durations and behavioural response patterns in conjunction with early stopping status. Finally, we conduct a comparison of group analyses between EPT versus healthy controls, to assess the effects of early stopping in this context.

2.0 Methodology

2.1 fMRI Analysis Methods

2.1.1 General linear model

Briefly, the general linear model involves convoluting a double gamma hemodynamic response function (HRF) with task indicator variables that denote timing of administration to reflect expected task-related BOLD responses. Voxel-level task-related regression parameters are estimated and represent the association of the observed response to expected task-activated BOLD signal. Thus, activation is assessed through statistical inference on regression parameters. For a given voxel up to time t (i.e. for scans 1 through t), the GLM takes the form:

$$Y_t = X_t B + E_t \qquad (1.1)$$

Where Y_t is a $t \times 1$ vector of observed BOLD signal intensities for the voxel up to time t, and E_t is a $t \times 1$ vector that represents the error components. X_t is a $t \times p$ design matrix and includes the

expected BOLD signal values per task. We also include cosine functions of increasing periodicity (scan duration*2, scan duration, scan duration/1.5, scan duration/2 and scan duration/2.5) to model physiological and other low frequency noise (22). For large periodicities, cosine functions are approximately linear for the time frame of scans we consider here, and hence are essentially collinear from a GLM modeling perspective. Five regressors were thus added to the design matrix. $B = [b_1 \dots b_j \dots b_p]^t$, a p × 1 regression coefficients vector. In this formulation, a regression parameter b_j can represent magnitude of association with task j. E_t is assumed to be distributed as multivariate normal with mean zero and covariance W_t, where W_t is a $t \times t$ matrix that represents the temporal autocorrelation structure. For spatial correlation, we conduct spatial smoothing, so do not explicitly model the spatial correlation structure. Y_t is assumed to have a multivariate normal probability distribution as follows:

$$f(Y_t, B, \sigma^2 W_t) = \frac{1}{(2\pi)^{t/2}|\sigma^2 W_t|} \exp\left(-\frac{1}{2}(Y_y - X_t B)'(\sigma^2 W_t)^{-1}(Y_t - X_t B)\right) \quad (1.2)$$

where $|\sigma^2 W_t|$ is the determinant of $\sigma^2 W_t$. Major sources of noise in fMRI data include brain metabolism, physiology, and spontaneous fluctuations (23).

We fit regression models in parallel for all voxels under consideration in a target region of interest (ROI), which could include the whole brain. Real-time analysis requires signal and image processing steps, as well as the continual updating of statistical estimates as new scan data are received from the scanner. Hence, given the large number of voxels to be analyzed, real-time fMRI presents “big data” computational challenges.

2.1.2 Sandwich Estimator

In our previous work (16), we assumed serial independence for computational simplicity. Here we recognize serial covariance using the sandwich estimator. The sandwich estimator is a robust, model-free covariance estimator that does not require distributional assumptions. Importantly, it still provides asymptotically consistent covariance matrix estimates, although convergence rates can be slow (20, 21). The approach is evaluated for computational feasibility for real time analysis, as well as plausibility in effectively modeling the error structure of the BOLD data over time within a voxel.

2.1.3 Wald’s Sequential Probability Ratio Test

At the voxel level, we can use the sequential analytic framework of (16, 24-28), to adaptively assess activation status using real-time fMRI. As we will demonstrate, Wald’s SPRT test statistic can serve as the basis of an efficient, sequential testing approach that can greatly reduce the need for experimental block administrations compared with fixed designs while attaining similar classification performance in simulation, and activation patterns with subject data. This approach relies on a SPRT statistic to conduct hypothesis testing, with the null hypothesis representing no activation with respect to a task, and the alternative hypothesis representing some threshold of

activation, as represented by a GLM parameter value (16). This statistic is updated with each new observation, and its value is compared with thresholds for stopping.

The general procedure of Wald's SPRT is described as follows. Consider a one-sided hypothesis of the form $H_0: c'\beta = c'\beta_0$ versus $H_a: c'\beta \geq c'\beta_1$, where $c'(\beta_1 - \beta_0) \geq 0$. Two-sided formulations are described in (24) and (16). Implementation of Wald's SPRT involves updating Wald's likelihood ratio statistic as new data are observed (24):

$$\Lambda_t = log\left(\frac{f\left(Y_t|c'\beta_1, \widehat{var}\left[c\hat{\beta}\right]\right)}{f\left(Y_t|c'\beta_0, \widehat{var}\left[c\hat{\beta}\right]\right)}\right) \qquad (1.3)$$

where $f(Y|c'\beta_0, \widehat{var}\left[c\hat{\beta}\right])$ and $f(Y|c'\beta_1, \widehat{var}\left[c\hat{\beta}\right])$ are the respective probability densities functions of Y_t given $c'\beta_0$ or $c'\beta_1$ is the true value of parameter of interest and conditioning on the estimated covariance. After Y_t is observed at a time point, t, one of three possible decisions is made according to the following rules:

1. Continue sampling if $B < \Lambda_t < A$
2. Stop sampling and accept H_0 if $\Lambda_t < B$
3. Stop sampling and accept H_a if $A < \Lambda_t$

where stopping boundaries (A, B) = (log((1-β_E)/α_E), log(β_E/(1-α_E))), and the target Type I and Type II error levels are respectively denoted as α_E and β_E. These error levels are specified before testing. Note that both the Type I and Type II error levels are controlled for with SPRT, as opposed to standard hypothesis test formulations that only control for Type I error level. Multiple SPRTs are conducted concurrently across voxels and boundary error levels can be adjusted for instance by Bonferroni correction to account for this simultaneous testing.

A practical modification of the original SPRT formulation for stopping is to consider the truncated SPRT (29), which will additionally call for stopping if an upper bound for the number of observations is reached. In our case, this is reached when a fixed number of blocks have been administered. Additional modifications include conducting two-stage estimation to allow sufficient observation for preliminary estimates of the voxel-level error variance from a first stage (30). With this estimate, we can derive an alternative hypothesis value for a linear contrast of task parameters $c'\beta$ that will correspond to a desired z-statistic value. As an illustration, suppose a z-statistic value of 3.10 is selected, as will be done below in our studies. Note 3.10 is the one-sided p-value = 0.001 - critical value for the standard normal distribution. Given an estimated value $\widehat{\sigma_t}^2$ from a first stage of length t scans, we solve for the value of $\theta_t = c'\beta$ that satisfies $\frac{\theta_t}{\sqrt{\widehat{var}[c'\beta]}} = 3.10$, where X_t is the design matrix up to scan t. This value becomes the alternative hypothesis, and it represents the voxel-level targeted activation magnitude threshold. We update the value of θ_t and $\widehat{\sigma^2}$at each scan, so that the alternative hypothesis is actually dynamic, since the estimation variance for $c'\hat{\beta}$ changes as well. In this framework, we use the sandwich estimator to estimate this variance.

Ultimately, we aggregate the findings of the voxel-level SPRTs to determine whether or not experimentation within a block design should be terminated early. A "global" stopping rule that considers all voxels in a region of interest (can be whole brain or smaller ROIs) that we have adopted is to terminate task administration when a predetermined percentage of voxels have been classified by their respective SPRTs. For instance, we have used 80% as a global stopping criterion. Note that 80% classified means either as active or non-active. We choose this cut-off as it is fairly strict, and as we will see, it will facilitate correspondence with full scan data results. We also choose Type I and Type II error levels that are relatively more stringent for Type I error. For the future, we are considering adaptive and more flexible, dynamically determined stopping rules beyond predetermined percentages. However, these will rely on more powerful computational resources than available for our current experiments. For now, we stop early only in a conservative manner. Note that for $c'\beta$ parameter values that are "in-between" the null and alternative hypothesis values, the SPRT is indifferent to preferring one hypothesis over the other. This leads to larger numbers of scans needed before a stopping boundary is crossed. So, we have to accept a lack of decisive stopping decisions for these cases in order for overall experimentation to stop early, even as θ_t decreases as t increases. This can be an acceptable trade-off for shorter experimental scan times and the ability to tailor experimentation.

2.2 MRI Data Collection Methods

Twelve healthy subjects were recruited, 7 males. They were aged 15-16 years old and 11 were right-handed. They had no known neurological conditions and a normal developmental history. A group of 11 adolescents born EPT were also recruited, 1 male. EPT is defined as being born at < 26-week gestation and weighing < 1000g. All were aged 15-17 years old and 8 were right-handed, 2 left-handed and 1 ambidextrous. All subjects were recruited as part of a larger study to evaluate functional and structural differences associated with mathematical abilities and working memory between those born EPT and those born at normal term. Adolescents were recruited as they can handle the stress of fMRI experimentation, are mathematically advanced enough and have had time to master the subject area. EPT subjects were included to show that differences with patient populations are detectable with our methods. A subsection of the full study is reported here to demonstrate the real-time analysis.

The subjects made one two-hour visit to the MRI department at University Hospitals Cleveland Medical Center (UHCMC). Ethics approval was obtained from the UHCMC Institutional Review Board office prior to the study and complied with the Declaration of Helsinki for human subject research. Subjects and their parents gave informed consent prior to taking part.

2.2.1 MRI protocols

The subjects were positioned head-first supine on the scanner bed with their head fixed in position using inflatable pads. An 8-channel head coil was used for data acquisition. Echo planar imaging scans were acquired on a Philips Ingenuity 3T PET/MR imager at UHCMC. The following fMRI scan parameters were used: TR = 3.0 s, TE = 35 ms, in-plane resolution was 1.797 mm^2

(matrix 128 x 128), slice thickness was 4 mm, number of slices = 36 slices and flip angle = 90°. A SENSE P reduction factor of 2 was implemented and scans were acquired in an ascending interleaved fashion.

In addition to the fMRI scans, a high-resolution T1-weighted anatomical image of the brain was also acquired. This was taken using a magnetic preparation gradient-echo sequence (3D IR TFE). Imaging parameters were: TR = 7.5 ms, TE = 3.7 ms, in-plane resolution was 1 mm^2 (matrix 256 x 256), slice thickness was 1 mm, number of slices = 200 slices and flip angle = 8°.

2.2.2 Stimulus protocols

During data acquisition subjects were presented with a mathematical version of the well-known 1-back memory task. It involved performing basic addition and subtraction calculations and required the answer to be remembered and compared to the next answer. Two difficulty levels were included. The protocol was developed by our lab as part of a battery to assess mathematical and working memory abilities in 14 – 17 year olds to evaluate the functional differences between those born EPT and those born at normal term. The stimulus was presented on an MRI compatible LCD monitor (manufactured by Cambridge Research Systems, Rochester, UK) positioned at the end of the bore and viewed via a mirror attached to the head coil. Equations were presented, for example, the subject may see "2 + 3 = ?". The subject was required to work out the answer and then remember it while working out the next equation, for example "1 + 4 = ?". If they thought the answers matched, then the subject pressed a button on a response box held in their right hand. If they thought the answers did not match, then they did nothing but remember the new answer to compare to the answer of the next equation. An example sequence is shown in Figure 1A.

The stimulus was presented in a block design, see Figure 1B and Table 1. Eight equations were presented per block. Each block lasted 36 seconds followed by 21 seconds of rest condition (fixation dot). Two difficulty levels were presented. The easier level consisted of single digit numbers to add or subtract and the answers were always a single digit. The harder level involved addition or subtraction of single or two-digit numbers and the answers were always two digits. Blocks of difficulty levels were alternated during the scan and a total of 6 blocks per level were presented. Note: although only 2 difficulty levels are used here, the setup is able to accommodate any number of difficulty levels. The full duration of the task was 238 scans or 11 minutes and 54 seconds. This was based on a moderate length of experimentation for a 1-back block design (e.g. see (31-34)), allowing approximately 6 minutes for each difficulty level.The visual stimulus was presented using an in-house custom written program that was developed using the Python programming language (Python Software Foundation, https://www.python.org/) and libraries from Psychopy - an open source visual presentation program (35-37). The program connected to a Cedrus Lumina controller to receive stimulus responses from the subject and trigger pulses from the MRI scanner (outputted every dynamic). The timing of the presentation of the visual stimulus was synchronized to the trigger pulses to ensure that stimulus images were displayed at the expected time. A Supervisor Window displayed on the experimenter's computer screen allowed the visual stimulus to be tracked throughout. It displayed the current block number being

presented, how many remaining blocks there were and when the subject responded. The program was also able to terminate one or both of the difficulty levels if it received a signal indicating the relevant areas in the fMRI data were sufficiently classified across voxels. The software is freely available from the Bitbucket repository: https://bitbucket.org/tatsuoka-lab/fmri-presentation.

2.2.3 Real-time fMRI acquisition

Real-time image transfer was achieved by XTC (eXTernal Control). This is a program integrated into the Philips scanner software and enabled by a research clinical science key. XTC communicates with the reconstruction and scanner processes on the scanner computer and interfaces to a network Client application using a minimalistic CORBA (Common Object Request Broker Architecture) (38) interface which uses TCP/IP as the transport layer. CORBA is platform independent, reliable, and has the ability to process large amounts of data with minimum overhead. Each CORBA message consisted of a hierarchical attribute collection identified with UUIDs (universally unique identifiers) (39). Messages carried reconstructed image data and meta-data containing details of scan protocols. Due to hospital network security protocols the reconstructed images were placed in a folder on the scanner computer and then pushed across to a Linux computer. To achieve necessary image transfer speeds to the scanner computer folder a modification to XTC was installed on the scanner to disable two-way communications as only one-way image transfer functionality was required. However, XTC does support two-way communication between the scanner and the Client.

The Linux computer was a custom-built server equipped with a solid state hard drive and two 8-core Intel Xeon E5-2687W processors running at 3.1 GHz and providing 40 MB L3 cache. It was installed with Centos 7.4 operating system. As the scans were received, custom written Python and Bash scripts implemented the analysis using core-based parallelization to preprocess the data and perform the SPRT statistical analysis. Preprocessing was performed using standard modules from AFNI (Analysis of Functional NeuroImages, https://afni.nimh.nih.gov) and FSL (FMRIB's Software Library, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). The analysis sequence is detailed in the following section. The setup is shown in Figure 2.

2.2.4 MRI preprocessing

At the beginning of the scanning session a single fMRI scan (3 sec) was acquired and used for coregistration (motion correction) purposes. In preparation, the skull was removed using FSL's Brain Extraction Tool (BET) (40) and a mask of the full brain was created. During the real-time adaptive fMRI scan session, new scans arrived every 3 seconds and were dumped in a folder on the Linux workstation where the following actions were applied to each one. AFNI's 'dcm2niix_afni' command was used to convert the .par/.rec files to nifti. Motion correction was performed using coregistration techniques. Every fMRI scan was realigned to the initial scan that was acquired before the task began, and AFNI's '3dvolreg' command was used. Spatial smoothing was also applied using an 8 mm kernel with AFNI's '3dmerge' command. The full brain mask created at the beginning of the session was applied using FSL's 'fslmaths' command to remove

noisy voxels outside the brain (voxels of no interest). The resulting images were then converted to ascii format for statistical analysis with SPRT.

2.2.5 fMRI SPRT analysis

The SPRT analysis was applied using highly-optimized C++ program that used Intel Cilk Plus library for multicore and vector processing of data. BLAS routines from Intel MKL were used to enable instruction-based acceleration for matrix computation. They are available from the Bitbucket repository at https://bitbucket.org/tatsuoka-lab. The design matrix was created prior to the scan session using AFNI's '3dDeconvolve' command to model the stimulus and HRF. It is possible to include the temporal derivatives of the HRF or other regressors in the design matrix where applicable in studies. Temporal derivatives were not included here due to the long durations of the block design used to present the task. Statistical analysis included the modeling of low frequency physiological noise and the associated removal of serial correlation using discrete cosine transforms. Motion parameters are also frequently used as regressors to remove correlated activations produced by movement. Here motion parameter regressors were not included with the estimation of the discrete cosine transforms due to the time limitation of conducting the analysis on each scan within the 3 second TR. The automatic determination of when to terminate the scanning is based on the Type I and Type II errors, α_E and β_E, as described above. A percentage of voxels that must be classified before termination was also specified during the setup, such as 80%. We also tested 2 scenarios using either 2-blocks or 4-blocks of easy and hard stimuli first stage administration before allowing for early stopping to occur.

3.0 Results

3.1 Individual Subject Results of SPRT

All subjects were able to complete the 1-back task. The median control subject response time across both difficulty levels was 1.44 sec (SD 0.51 sec), and median task accuracy was 90.8 % (SD 20.2 %). When these are broken down by difficulty level, the easy level median task accuracy was 86.1 % (SD 22.6 %) with median response time of 1.28 sec (SD 0.54 sec); and the hard level median task accuracy was 90.0 % (SD 18.4 %) with median response time of 1.56 sec (SD 0.51 sec). EPT subjects had a slightly longer overall median response time of 1.91 sec (SD 0.48 sec) and overall median task accuracy was lower at 65.8 % (SD 21.2 %). For the easy level, the median accuracy was 72.2 % (SD 24.2 %) and median response time was 1.63 sec (SD 0.49 sec). For the hard level the median accuracy was 70.0 % (SD 19.8 %) with a median response time of 2.10 sec (SD 0.54 sec).

Real-time transfer speeds between the scanner and the Linux computer were consistently fast, with individual scan files taking less than 150 milliseconds to transfer. All subject scans were processed within the 3 second TR period. Offline testing showed that the subject with the largest number of voxels (subject 21 with 135,379 voxels) was processed in just 5 minutes and 45 seconds. The subject with the fewest number of voxels (subject 14 with 77,359 voxels) was processed in 5 minutes and 2 seconds. Thus, it is feasible for a TR of 2 seconds to be used with the software, depending on transfer speeds.

Inspection of the z-score maps for each subject showed that generally, across subjects, the largest activations were centered bilaterally around the inferior and superior parietal areas, taking in the intraparietal sulcus, a region highly associated with mathematical functioning. Further activations were seen in the cuneus. These are most likely correlated with the visual processing associated with the task. Additional activations were seen in the precuneus, bilateral areas in the medial frontal gyrus, anterior cingulate, insula and inferior frontal gyrus. These areas are often associated with attention and memory systems (41, 42).

Functional MRI data were analyzed using typical Type I and II error levels, to test stopping time performance, with $\alpha_E = 0.001$, $\beta_E = 0.1$ (43, 44). The results of the SPRT versions are presented in Tables 2 A-D. A full-length task protocol comprised 238 scans and lasted 11 minutes and 54 seconds. Where 2-blocks per difficulty level of stimulus administration were used before allowing for early stopping, the first 78 scans were used for the first stage of experimentation. Where 4-blocks were used, 154 scans were used for the first stage.

The stopping times for the 2- and 4-block first stage lengths are reported in Table 2. Stopping was reached at 80% of voxels classified as either active or non-active in around 54% of cases in both scenarios for both difficulty levels. At 80% classification for control subjects, 58% of subjects stopped early for the easy level with both the 2-block and 4-block first stages. For the hard level, 58% with 2-blocks and 42% with 4-blocks stopped early. For EPT subjects, 55% of subjects stopped early for the easy level for both 2- and 4-block first stage conditions. For the hard level, 45% of subjects using 4-blocks as a minimum still stopped early and 64% of subjects using 2-blocks as a minimum stopped early. The median stopping duration for both difficulty levels for control subjects was 3 blocks of easy and 2 blocks of hard stimulus administration for 2-blocks first stage. For 4-blocks first stage, the median stopping time was 5 easy, 4 hard for the easy level and 4 easy, 4 hard for the hard level. In EPT subjects, the median stopping time for 2-block first stage was 2 easy and 2 hard blocks of stimuli. For 4-blocks first stage, the median stopping time was 4 easy and 4 hard for both difficulty levels. Depending on the number of first stage blocks, time savings of 1/3 to 2/3 can be achieved.

An early stopping rule based on a classification rate of at least 90% was also tested. There were very few instances when early stopping occurred. For the 2-blocks first stage condition, 3 subjects stopped early for the easy level and 1 subject for the hard level. Only 1 subject stopped early under the 4-blocks first stage condition for the easy level.

The activation maps under the different conditions are shown in Figure 3 for a sample subject (subject 9). For the 2-block first stage, this subject terminated after 2 blocks of easy and hard administration for the easy level (scan 79) and after 3 blocks of easy and 2 blocks of hard administration for the hard level (scan 97). For the 4-block minimum, this subject terminated at scan 155, equal to 4 blocks of easy and hard stimulus administration, for both difficulty levels. The images show that at scan 79 there is very little activity present and the majority of the voxel classifications are non-active. By scan 155, there is much more activity which has a similar pattern to the full duration result (scan 238). The extent is not quite as large as scan 238, however the

foci of the clusters do overlap between the two time points. As mentioned, this is likely due to the alternative hypothesis threshold $c'\beta$ value corresponding to relatively lower z-score values as the number of scans increase. This pattern of 'growing' activations for given alternative hypothesis $c'\beta$ over scan duration is thus typical of our early stopping data, particularly for the 2-block initial stage. Also shown in Figure 3 is the full duration scan threshold at z = 4.0. Visual inspection of the z-score maps at the stopping scan for other subjects revealed similar patterns. In most instances, the additional active voxels at full duration were around the edges of existing clusters at the early stopping scans. Further images of other subjects are presented in Table S1 in the Supplementary Information document. Plots of the percentage of active and non active voxels classified at each scan are given in Table S2 of the same document. The overlaps between early stopping and full duration maps are also explored further in the right-hand columns of Table 2 where we show the number of active voxels in common spatially between the two durations. Although some of these show less than 50% overlap with scan 238, it can be seen that this is due to the smaller cluster sizes with early stopping. The median spatial overlap where early stopping occurs for control subjects was 27.9% (SD 30.2%) for the easy level, 2-blocks and 68.5% (SD 15.4%) for the easy level, 4-blocks. For the hard level, there was 26.0% (SD 19.9%) and 44.6% (SD 21.3%) for the 2-block and 4-block first stages, respectively. For EPT subjects the median overlap was 34.2% (SD 34.0%) and 33.5% (SD 34.3%) for the easy level, 2- and 4-block first stages, respectively. For the hard level, the median overlap values were 14.6% (SD 9.4%) and 77.6% (SD 25.2%) for 2-block and 4-block first stages.

In summary, although there are similar rates of early termination between the 2-block and 4-block first stage cases, the detected activation patterns suggest that using 4-blocks of stimulus administration is more suited to determining active voxels and produces an image most similar to full duration scans. In Figure 4, to illustrate the rate of decrease of the estimated σ_t^2 values, we present a plot of sandwich estimator $\widehat{\sigma_t}$ values for one subject across a set of voxels over the duration of the experiment. These values give an indication that stopping based on the theta values at the end of the 2-block first stage may be too early for correspondence with full duration scans, as the estimated standard deviations are relatively larger. This implies that the alternative hypotheses can have much larger theta values early on compared to full duration, while this difference in theta values is less after 4-blocks. Note that in some subjects there is some volatility due to subject movement. Below, we consider full duration z-score thresholds of 4.0 versus early stopping results with a threshold of 3.1, so that activation magnitudes considered as active are more comparable.

In all but three EPT subjects the active voxel count increases with scan duration. Subjects 13, 15 and 17 are the exception. Subject 13 demonstrates very few active voxels at all and there is almost no consistency in location. Further investigation shows large relative framewise displacement occurs frequently throughout the scan and many of the responses have been missed or have relatively long response times, 60.5% correct overall and 2.08 s (SD 0.97 s) average response time (see plots for subject 13 in Table S3 of Supplementary Information document). Taken together these suggest that either the task level may not have been aimed at the right level and/or the subject may have been uncomfortable and distracted in the scanner thereby attending to the task less than required for robust activations to occur. Subject 15

demonstrates cluster sizes that decrease over time. Framewise displacement shows very little motion, particularly from scan 180 onwards. The response plots (in Supplementary Information) show the subject is paying attention and responding. Subject 17 has a similar pattern of decreasing cluster sizes. The framewise displacement plots indicate a moderate amount of motion throughout. Although the subject has missed many of the task questions (65.8% correct), the pattern of responding indicates they are awake and attending to the task.

In general, EPT subjects demonstrated more motion. The median number of scans with framewise displacement above a threshold of 0.9 mm, threshold determined from (45), was 5 scans (SD 43 scans) for EPT subjects and 2.5 scans (SD 8 scans) for control subjects. One EPT subject passed the threshold a total of 124 scans out of 238 scans. In contrast, the control subject with the maximum number of threshold passes was 30/238 scans. This is further demonstrated in Figure 5 where we show subject counts for each scan when the threshold has been passed. For both EPT and control subjects, it is clear that subjects are moving more frequently in the second half of the scans and supports stopping early to reduce motion artifacts and noise in the data. Also, those born EPT can have structural abnormalities of the brain which can affect fMRI results and 2 subjects reported here had clear abnormalities that were obvious in this low resolution data. Less obvious abnormalities may have been present in the other subjects.

3.2 Group Analysis Results

The preprocessed files from the SPRT analysis can be used directly to perform a group analysis using AFNI's 3dMEMA command (Mixed Effects Meta Analysis tool) (46). However, a group analysis was carried out using FSL which instead merges all subject data to conduct a combined mixed models analysis. We demonstrate that the data collected in real-time can still be used in a typical post-hoc analysis. Raw data was preprocessed with FSL FEAT (47). Motion correction was performed using a rigid body transform, spatial smoothing with a full-width-at-half-maximum Gaussian kernel of 6 mm was applied, high pass temporal filtering of 90 s was carried out and coregistration to (MNI) standard space was done before performing a first level individual GLM analysis. The statistical output from these were used to perform the higher level group statistics using FLAME 1 (FMRIB's Local Analysis of Mixed Effects, (48)). The results for the 1-back easy and hard contrasts for the 2- and 4-block first stage conditions for EPT and control subjects are shown in Figure 6. Location of activity is listed in Table S4 of the Supplementary Information. The group results of full scan durations are compared to the group results using only the scans up to the early stopping point for each subject for each difficulty level and number of blocks completed before early stopping was allowed.

For both subject groups, there is consistency in location and overlap of clusters between the group activations for each difficulty level and 2- or 4-block first stages. The control subjects show strong activations in the anterior cingulate and bilateral parietal regions, see Tables 4 and S4, and Figure 6. This supports that the experiments are not being stopped too early.

The EPT subjects demonstrate more right sided activity and smaller cluster sizes by comparison to control subjects across all stopping points. In order to understand this result it is necessary to

consider neuropsychological skills and structural and functional brain changes within the group. Behavioral data collected as part of our wider study shows that nearly two thirds (63.6 %) of the EPT cohort have lower working memory function, compared to just over one third (35.7 %) of controls subjects. Working memory is a key skill required for both mathematics and this numerical 1-back task. Recall the lower accuracy and longer response times in the EPT group. fMRI studies on dyscalculia (difficulty in learning and performing mathematics) suggest that there is greater heterogeneity in activations with a more diffuse pattern being apparent (49, 50). Additionally, there is overlap in structural differences in white matter integrity, as measured from diffusion weighted imaging studies, between those born EPT and those with dyscalculia including inferior fronto-occipital fasciculus and the inferior and superior longitudinal fasciculi (51-54). These connect crucial areas associated with mathematics and working memory. A more diffuse and variable pattern of functional activity, perhaps partly due to structural differences, may confound a group analysis in this instance. More data points from individuals do seem to improve the results, perhaps allowing the variability to converge somewhat. This is supported by the change in variance for the group between early stopping with 2- and 4-block first stages and full duration analyses, see right-hand column of Table 4. The control group variances are much lower throughout. For both groups, reducing the z-score threshold of the early stopping maps to z = 2.58 produces maps similar to the full duration maps indicating the trend towards the full duration results (not shown).

4.0 Discussion

We present a workflow for the implementation of an adaptive real-time fMRI system that allows for statistically-driven dynamic adjustment of experimentation based on voxel-level SPRT. We also explored imposing 2 different first stage lengths before early stopping using either 2- or 4-blocks of easy and hard stimulus administration. We show that this dynamic and adaptive statistical approach is comparable to corresponding fixed experimental designs in terms of detected activation, particularly under the 4-block condition and when adjusting for stricter z-score thresholds for full scan to account for reduced estimation variance. At the same time, time savings in experiment durations ranged between approximately one third and two thirds of the full protocol. The SPRT results in Table 2 indicate that results are fairly comparable with a full protocol analysis for the 4-block first stage condition. The 2-block first stage often led to most voxels being classified as non-active. The statistical stopping criteria of the approach helped ensure accurate characterizations of activation patterns while reducing time needed for scanning and providing a framework for flexible yet systematic dynamic paradigms.

Our SPRT approach was effective at detecting brain activity at the individual level with early stopping in both the control and EPT groups. We note the individual variability among subjects in early stopping performance. Factors that can affect stopping times include the magnitudes of activation, variability in task performance and subject attention, subject motion and the noise levels in the BOLD signal. In the future, we will consider development of adaptive approaches to individualize global stop rule threshold selection that can encourage early stopping while balancing the potential for improved activation classification with continued experimentation. Although not considered here, these BOLD signal-based early stopping rules could also possibly

be enriched by incorporating individual motion displacement patterns, as well as behavioural measures such as correctness rates in experimentation.

There are many possible applications in the research setting where individual level results may be the focus. A possible clinical application may be in clinical assessments for presurgical evaluation for brain surgery in patients with brain tumors or epilepsy. A group analysis was also feasible using the early stopping data in controls. A possible limitation was discovered in performing a group analysis of the EPT group who demonstrated greater variability in location at the individual level. While it is feasible to apply our approach for patient group studies, consideration should be given to the particular patient group of interest and the likely differences in brain activity when making the decision to stop early. If early stopping is attractive, such as to reduce burden for an impaired population, we conjecture that larger sample sizes or stricter early stopping criteria may help overcome this larger variability.

In the modified SPRT framework, other α_E, β_{E-} pairs were considered as well, to test how different combinations impact activity detection and early stopping. For instance, given selection of α_E = 0.001 and β_E = 0.01, overall stopping did not occur. In this case, the more stringent choice of β_E makes it more difficult to cross either of the SPRT thresholds, although here the active voxels are still classified relatively quickly. We also saw that for α_E = 0.001 and β_E = 0.1, we gain dramatic gains in shorter scan durations for experimentation, across both of the experimental conditions that are considered. Note, that for SPRT, stopping is less likely for values in-between the null and alternative hypothesis values. As we selected a stringent alternative value (with z-statistic approximately equal to 3.1), this makes stopping less likely if we focus on reducing Type II error. Also, since we are stopping "globally", voxel-level SPRT values exceed stopping thresholds in many cases.

Note that it also is important to balance the stopping times for active and non-active decisions, so that there is not a systematic bias towards one activation status being classified faster and driving attainment of the global stopping rule, while the other status has less decisive classification. For instance, a potential concern is if the non-active voxels do not require as many scans to satisfy the SPRT criterion. If there also is a high percentage of non-actives in the ROI for the experiment, then the global stopping rule would result in decisively classifying only non-active voxels. On the other hand, if the SPRT stopping times for active and non-active voxels are generally similar, then there will be a proportional balance of decisively classified active and non-active voxels. This mix can be adjusted by selection of the respective error thresholds that determine stopping boundaries. We show this balance in plots of the percentages of active and non-active voxels classified throughout the scan for each subject in Table S2 of the Supplementary Information document.

In terms of the global stop rule threshold, we observed that for the cases under consideration, stopping when 80% of voxels in the full brain (or smaller ROI) respectively satisfy their SPRT-based stopping criterion generally leads to early stopping of stimulus administration, while also leading to comparable activation classification as with the full protocol, after t-test score threshold adjustment. The stricter 90% criterion was infrequently satisfied, and did not often

lead to early stopping of experimentation. Recall that when GLM parameter values are "in-between" the null and alternative hypothesis values, SPRT-based stopping is less likely at the voxel level. This does become less of an issue, since θ_t, the alternative hypothesis threshold for $c'\beta$, decreases in value as more scans are accrued. A 100% stopping rule is thus not feasible. The 80% rule seems conservative in that not all participants are stopped early, but there are high levels of correspondence in individual and group level activation maps with full durations, particularly when the initial first stage is 4 blocks. The 4-block initial stage is justified by comparison of estimation variances with full scan durations and the high individual and group level correspondence in activation analyses. In contrast, for the 2-block initial stage, estimation variances are clearly higher, and correspondence to full scan results are lower. The 4-block initial stage provides less opportunity for efficiency gains, as the window for early stopping is narrower, but it is more prudent. It is possible that a 3-block initial stage could provide comparable results as the 4-block initial stage, but it was not explored here.

Here we demonstrated full brain analytics with parallelization using MKL Intel libraries for matrix computation with two Xeon E5-2687W 8-core processors. It is also feasible to consider only partial brain volumes where experiments demand more consideration of a particular area. Future directions for the study are to implement the SPRT and Bayesian sequential estimation methods using distributed computing approaches to increase processing speed allowing full brain real-time analyses and advance stopping rule methods in shorter scan times. This could include dynamic selection of global stopping rule thresholds, and look-ahead stopping rules that rely on probabilistic prediction.

5.0 Conclusion

We introduce a systematic, statistically-based approach to dynamic experimentation with real-time fMRI. Saving in scan time and accurate voxel activation detection can be achieved, while redundant experimentation in block designs is reduced. We investigate different aspects of how to determine early stopping rules. With fMRI data from an actual study with both extremely premature birth and normal birth adolescents, we see different impacts from early stopping on activation analyses. These analyses can be viewed as intended on a training sample to guide implementation of early stopping in future studies involving the same experiments and study populations. These methods lay a foundation for future dynamic experimentation approaches with fMRI, and for real-time quality control analyses to assess statistical error performance across voxels. Use of high performance computing will enable the advent of these more sophisticated real-time experimental designs.

Declaration of conflicts of interest: All authors declare no conflicts of interest.

Author contributions

SC – Study design, analysis and interpretation of data, drafting of manuscript

WC – Study design, analysis and interpretation of data, software development, drafting of manuscript
JF – Study design and software development
HF – Study design and technical support, drafting of manuscript
JSG – Software development
JZ – Analysis and interpretation of data
CT – Study design, analysis and interpretation of data, drafting of manuscript

Funding: This study was supported by Phillips Healthcare and the National Science Foundation (Award number: 1561716).

Acknowledgements
None

Data availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

1. A. A. Alegria, M. Wulff, H. Brinson, G. J. Barker, L. J. Norman, D. Brandeis, D. Stahl, A. S. David, E. Taylor, V. Giampietro and K. Rubia: Real-Time fMRI Neurofeedback in Adolescents with Attention Deficit Hyperactivity Disorder. *Human Brain Mapping*, 38(6), 3190-3209 (2017) doi:10.1002/hbm.23584
2. M. S. Sherwood, J. H. Kane, M. P. Weisend and J. G. Parker: Enhanced control of dorsolateral prefrontal cortex neurophysiology with real-time functional magnetic resonance imaging (rt-fMRI) neurofeedback training and working memory practice. *Neuroimage*, 124, 214-223 (2016) doi:10.1016/j.neuroimage.2015.08.074
3. E. J. Lawrence, L. Su, G. J. Barker, N. Medford, J. Dalton, S. C. R. Williams, N. Birbaumer, R. Veit, S. Ranganatha, J. Bodurka, M. Brammer, V. Giampietro and A. S. David: Self-regulation of the anterior insula: Reinforcement learning using real-time fMRI neurofeedback. *Neuroimage*, 88, 113-124 (2014) doi:10.1016/j.neuroimage.2013.10.069
4. C. Hohenfeld, N. Nellessen, I. Dogan, H. Kuhn, C. Muller, F. Papa, S. Ketteler, R. Goebel, A. Heinecke, N. J. Shah, J. B. Schulz, M. Reske and K. Reetz: Cognitive Improvement and Brain Changes after Real-Time Functional MRI Neurofeedback Training in Healthy Elderly and Prodromal Alzheimer's Disease. *Frontiers in Neurology*, 8 (2017) doi:10.3389/fneur.2017.00384
5. L. Subramanian, J. V. Hindle, S. Johnston, M. V. Roberts, M. Husain, R. Goebel and D. Linden: Real-Time Functional Magnetic Resonance Imaging Neurofeedback for Treatment of Parkinson's Disease. *Journal of Neuroscience*, 31(45), 16309-16317 (2011) doi:10.1523/jneurosci.3498-11.2011
6. L. Subramanian, M. B. Morris, M. Brosnan, D. L. Turner, H. R. Morris and D. E. J. Linden: Functional Magnetic Resonance Imaging Neurofeedback-guided Motor Imagery Training and

Motor Training for Parkinson's Disease: Randomized Trial. *Frontiers in Behavioral Neuroscience*, 10 (2016) doi:10.3389/fnbeh.2016.00111
7. S. Lui, X. H. J. Zhou, J. A. Sweeney and Q. Y. Gong: Psychoradiology: The Frontier of Neuroimaging in Psychiatry. *Radiology*, 281(2), 357-372 (2016) doi:10.1148/radiol.2016152149
8. R. W. Cox, A. Jesmanowicz and J. S. Hyde: REAL-TIME FUNCTIONAL MAGNETIC-RESONANCE-IMAGING. *Magnetic Resonance in Medicine*, 33(2), 230-236 (1995) doi:10.1002/mrm.1910330213
9. R. Goebel: BrainVoyager - Past, present, future. *Neuroimage*, 62(2), 748-756 (2012) doi:10.1016/j.neuroimage.2012.01.083
10. J. R. Sato, R. Basilio, F. F. Paiva, G. J. Garrido, I. E. Bramati, P. Bado, F. Tovar-Moll, R. Zahn and J. Moll: Real-Time fMRI Pattern Decoding and Neurofeedback Using FRIEND: An FSL-Integrated BCI Toolbox. *Plos One*, 8(12) (2013) doi:10.1371/journal.pone.0081658
11. J. J. Yoo, O. Hinds, N. Ofen, T. W. Thompson, S. Whitfield-Gabrieli, C. Triantafyllou and J. D. E. Gabrieli: When the brain is prepared to learn: Enhancing human learning using real-time fMRI. *Neuroimage*, 59(1), 846-852 (2012) doi:10.1016/j.neuroimage.2011.07.063
12. O. Hinds, S. Ghosh, T. W. Thompson, J. J. Yoo, S. Whitfield-Gabrieli, C. Triantafyllou and J. D. E. Gabrieli: Computing moment-to-moment BOLD activation for real-time neurofeedback. *Neuroimage*, 54(1), 361-368 (2011) doi:10.1016/j.neuroimage.2010.07.060
13. M. T. Debettencourt, J. D. Cohen, R. F. Lee, K. A. Norman and N. B. Turk-Browne: Closed-loop training of attention with real-time brain imaging. *Nature Neuroscience*, 18(3), 470-165 (2015) doi:10.1038/nn.3940
14. R. Lorenz, R. P. Monti, I. R. Violante, C. Anagnostopoulos, A. A. Faisal, G. Montana and R. Leech: The Automatic Neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI. *Neuroimage*, 129, 320-334 (2016) doi:10.1016/j.neuroimage.2016.01.032
15. R. Lorenz, I. R. Violante, R. P. Monti, G. Montana, A. Hampshire and R. Leech: Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization. *Nature Communications*, 9 (2018) doi:10.1038/s41467-018-03657-3
16. I. J. Feng, A. I. Jack and C. Tatsuoka: Dynamic Adjustment of Stimuli in Real Time Functional Magnetic Resonance Imaging. *Plos One*, 10(3) (2015) doi:10.1371/journal.pone.0117942
17. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich and S. M. Smith: FSL. *Neuroimage*, 62(2), 782-790 (2012) doi:10.1016/j.neuroimage.2011.09.015
18. N. Lazar: The Statistical Analysis of Functional MRI Data. Springer-Verlag New York, (2008) doi:10.1007/978-0-387-78191-4
19. R. W. Cox: AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3), 162-173 (1996) doi:10.1006/cbmr.1996.0014
20. R. J. Carroll, S. Wang, D. G. Simpson, A. J. Stromberg and D. Ruppert: The Sandwich (Robust Covariance Matrix) Estimator. In: Texas A&M University, (1998)
21. G. Kauermann and R. J. Carroll: A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456), 1387-1396 (2001) doi:10.1198/016214501753382309

22. K. J. Friston, O. Josephs, E. Zarahn, A. P. Holmes, S. Rouquette and J. B. Poline: To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *Neuroimage*, 12(2), 196-208 (2000) doi:10.1006/nimg.2000.0609
23. C. Yan, D. Liu, Y. He, Q. Zou, C. Zhu, X. Zuo, X. Long and Y. Zang: Spontaneous brain activity in the default mode network is sensitive to different resting-state conditions with limited cognitive load. *PLoS One*, 4(5), e5743 (2009) doi:10.1371/journal.pone.0005743
24. A. Wald: Sequential Analysis. Wiley, New York (1947)
25. A. Wald and J. Wolfowitz: OPTIMUM CHARACTER OF THE SEQUENTIAL PROBABILITY RATIO TEST. *Annals of Mathematical Statistics*, 19(3), 326-339 (1948) doi:10.1214/aoms/1177730197
26. D. R. Cox: Large Sample Sequential Tests for Composite Hypotheses. *The Indian Journal of Statistics, Series A (1961-2002)*, 25(1), 5-12 (1963)
27. S. Tantaratana and J. B. Thomas: TRUNCATED SEQUENTIAL PROBABILITY RATIO TEST. *Information Sciences*, 13(3), 283-300 (1977) doi:10.1016/0020-0255(77)90050-0
28. J. X. Li: Sequential probability ratio tests for generalized linear mixed models. In: University of California, Riverside, (2010)
29. J. B. T. Sawasd Tantaratana: Truncated sequential probability ratio test. *Information Sciences*, 13(3), 283-300 (1977) doi:http://dx.doi.org/10.1016/0020-0255(77)90050-0
30. W. J. Hall: SOME SEQUENTIAL ANALOGS OF STEINS 2-STAGE TEST. *Biometrika*, 49(3-4), 367-& (1962) doi:10.1093/biomet/49.3-4.367
31. X. Di, H. M. Zhang and B. B. Biswal: Anterior cingulate cortex differently modulates frontoparietal functional connectivity between resting-state and working memory tasks. *Human Brain Mapping*, 41(7), 1797-1805 (2020) doi:10.1002/hbm.24912
32. J. Z. Huang, L. Xie, R. W. Guo, J. H. Wang, J. Q. Lin, Z. B. Sun, S. X. Duan, Z. R. Lin, H. Li and S. H. Ma: Abnormal brain activity patterns during spatial working memory task in patients with end-stage renal disease on maintenance hemodialysis: a fMRI study. *Brain Imaging and Behavior* (2020) doi:10.1007/s11682-020-00383-7
33. M. Daamen, J. G. Bauml, L. Scheef, C. Sorg, B. Busch, N. Baumann, P. Bartmann, D. Wolke, A. Wohlschlager and H. Boecker: Working Memory in Preterm-Born Adults: Load-Dependent Compensatory Activity of the Posterior Default Mode Network. *Human Brain Mapping*, 36(3), 1121-1137 (2015) doi:10.1002/hbm.22691
34. G. R. Poudel, J. C. Stout, J. F. Dominguez, M. A. Gray, L. Salmon, A. Churchyard, P. Chua, B. Borowsky, G. F. Egan and N. Georgiou-Karistianis: Functional changes during working memory in Huntington's disease: 30-month longitudinal data from the IMAGE-HD study. *Brain Structure & Function*, 220(1), 501-512 (2015) doi:10.1007/s00429-013-0670-z
35. J. W. Peirce: PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8-13 (2007) doi:10.1016/j.jneumeth.2006.11.017
36. J. W. Peirce: Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2 (2009) doi:10.3389/neuro.11.010.2008
37. J. Peirce and M. MacAskill: Building Experiments in Pscychopy. Sage Publications Ltd, London, England (2018)
38. D. C. Schmidt, D. L. Levine and S. Mungee: The design of the TAO real-time object request broker. *Computer Communications*, 21(4), 294-324 (1998) doi:10.1016/s0140-3664(97)00165-5

39. ITU-T: Series X: Data Networks and Open System Communications. In: Ed I. T. S. S. o. ITU. ISO/IEC, (2005)
40. S. M. Smith: Fast robust automated brain extraction. *Human Brain Mapping*, 17(3), 143-155 (2002) doi:10.1002/hbm.10062
41. V. Menon: Memory and cognitive control circuits in mathematical cognition and learning. *Mathematical Brain across the Lifespan*, 227, 159-186 (2016) doi:10.1016/bs.pbr.2016.04.026
42. A. E. Cavanna and M. R. Trimble: The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129, 564-583 (2006) doi:10.1093/brain/awl004
43. H. R. Cremers, T. D. Wager and T. Yarkoni: The relation between statistical power and inference in fMRI. *Plos One*, 12(11) (2017) doi:10.1371/journal.pone.0184923
44. A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar and S. Chaudhury Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 18, 127-131 (2009) doi:0.4103/0972-6748.62274
45. J. S. Siegel, J. D. Power, J. W. Dubis, A. C. Vogel, J. A. Church, B. L. Schlaggar and S. E. Petersen: Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, 35(5), 1981-1996 (2014) doi:10.1002/hbm.22307
46. G. Chen, Z. S. Saad, A. R. Nath, M. S. Beauchamp and R. W. Cox: FMRI group analysis combining effect estimates and their variances. *Neuroimage*, 60(1), 747-765 (2012) doi:10.1016/j.neuroimage.2011.12.060
47. M. W. Woolrich, B. D. Ripley, M. Brady and S. M. Smith: Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, 14(6), 1370-1386 (2001) doi:10.1006/nimg.2001.0931
48. M. W. Woolrich, T. E. J. Behrens, C. F. Beckmann, M. Jenkinson and S. M. Smith: Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*, 21(4), 1732-1747 (2004) doi:10.1016/j.neuroimage.2003.12.023
49. K. Kucian, T. Loenneker, T. Dietrich, M. Dosch, E. Martin and M. von Aster: Impaired neural networks for approximate calculation in dyscalculic children: a functional MRI study. *Behav Brain Funct .* 2, 31 (2006) doi:10.1186/1744-9081-2-31
50. M. M. Berl, C. J. Vaidya and W. D. Gaillard: Functional imaging of developmental and adaptive changes in neurocognition. *Neuroimage*, 30(3), 679-691 (2006) doi:10.1016/j.neuroimage.2005.10.007
51. I. M. Loe, J. N. Adams and H. M. Feldman: Executive Function in Relation to White Matter in Preterm and Full Term Children. *Frontiers in Pediatrics*, 6 (2019) doi:10.3389/fped.2018.00418
52. J. M. Young, B. R. Morgan, H. E. A. Whyte, W. Lee, M. L. Smith, C. Raybaud, M. M. Shroff, J. G. Sled and M. J. Taylor: Longitudinal Study of White Matter Development and Outcomes in Children Born Very Preterm. *Cerebral Cortex*, 27(8), 4094-4105 (2017) doi:10.1093/cercor/bhw221
53. E. Rykhlevskaia, L. Q. Uddin, L. Kondos and V. Menon: Neuroanatomical correlates of developmental dyscalculia: combined evidence from morphometry and tractography. *Frontiers in Human Neuroscience*, 3 (2009) doi:10.3389/neuro.09.051.2009

54. K. Kucian, S. S. Ashkenazi, J. Hanggi, S. Rotzer, L. Jancke, E. Martin and M. von Aster: Developmental dyscalculia: a dysconnection syndrome? *Brain Structure & Function*, 219(5), 1721-1733 (2014) doi:10.1007/s00429-013-0597-4

Figure 1: A) Sample 1-back protocols demonstrating the two difficulty levels. B) Block design and timings of each difficulty level.

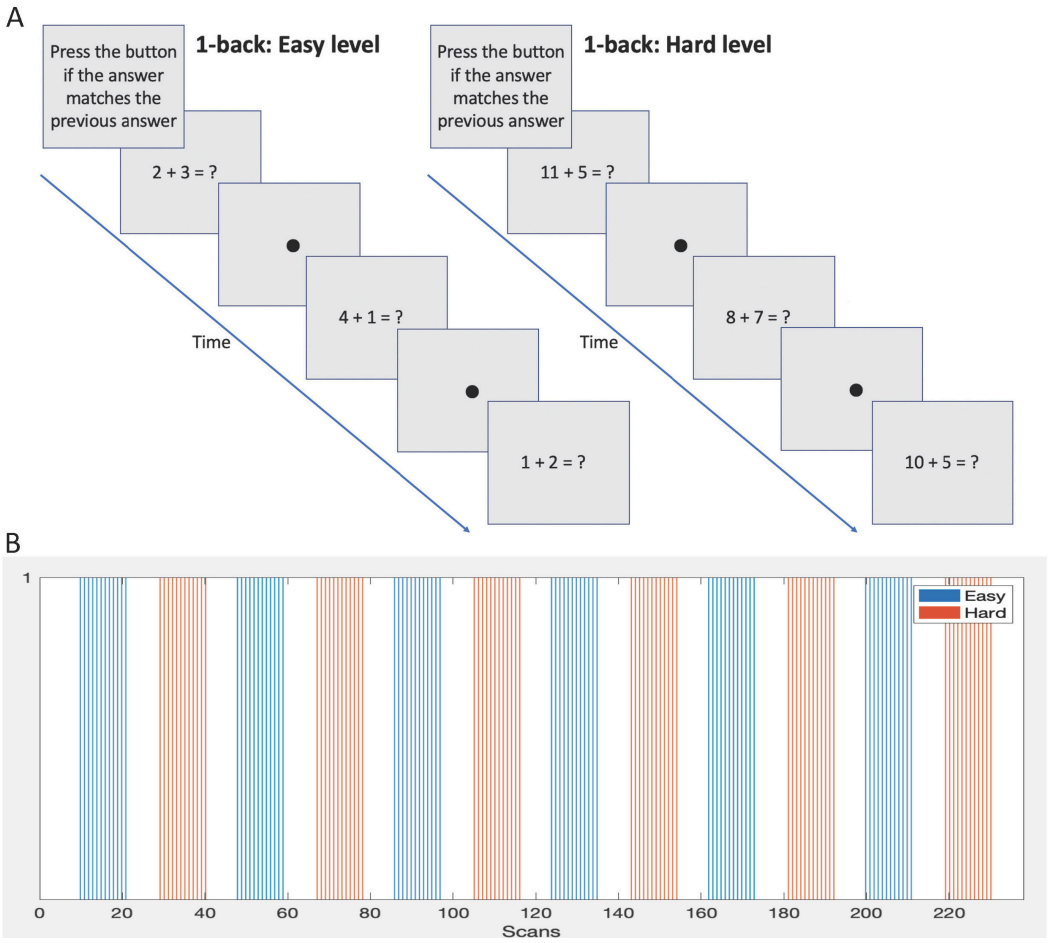

Figure 2: Schematic of the experimental setup of the dynamic real-time fMRI process. The equations were presented to the subject while the scans were acquired and exported to the Linux workstation for processing with SPRT statistics. The results were relayed back to the presentation program with instruction to either continue or terminate the stimulus.

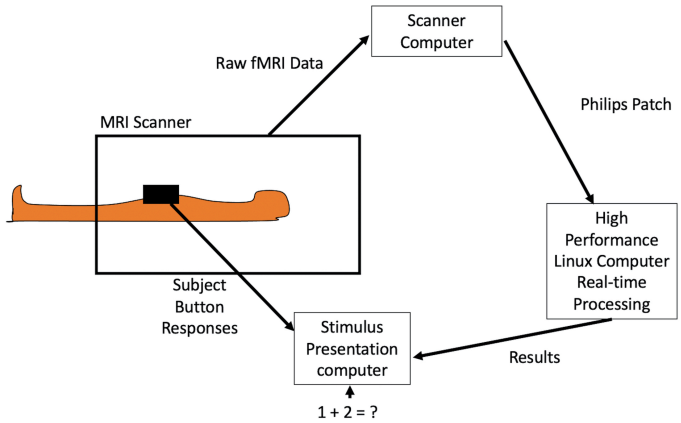

Figure 3: Full brain activation z-score maps that are associated with the easy (top) and hard (bottom) level 1-back task for 1 subject (number 9). Left in red shows SPRT results with early stopping determined using 2-blocks of stimulus administration for the first stage. Middle in green shows the early stopping activation maps with 4-blocks of stimulus administration first stage. Right in blue shows full duration results (activations at scan 238). $P \leq 0.001$ uncorrected, $z > 3.1$. Far right in yellow shows the full duration results with a higher threshold of $z > 4.0$. Shown in subject own space. R = right, L = left, A = anterior, P = posterior.

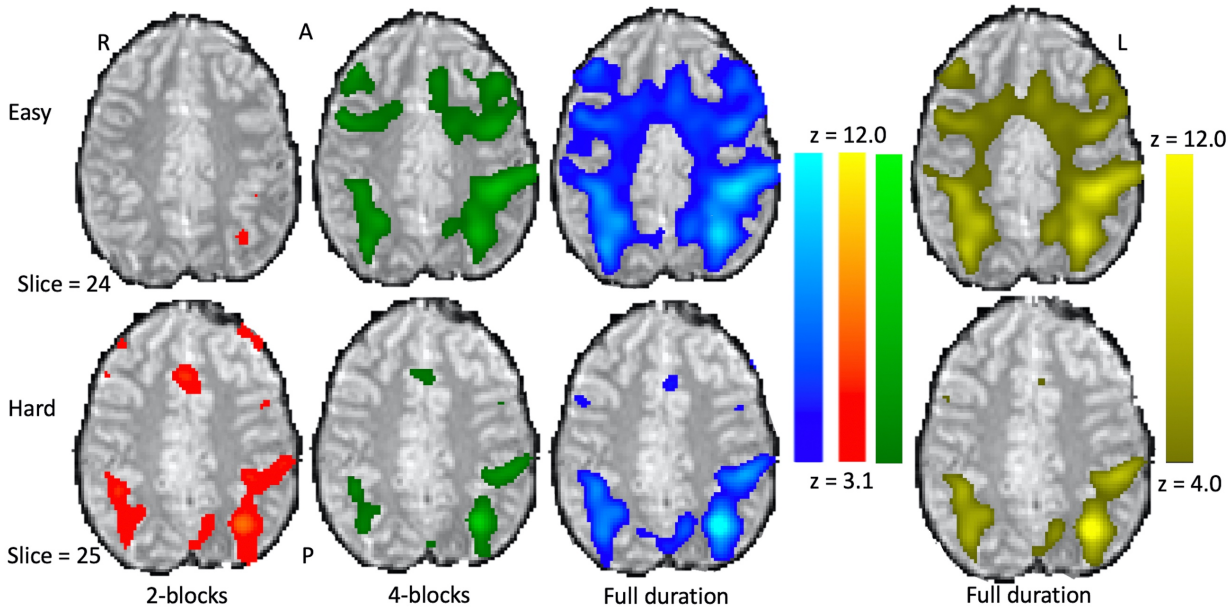

Figure 4: Estimated standard deviations for $c_1\widehat{\beta}$. Plots for 3 sample active (bottom) and non-active (top) voxels from a control subject (subject 3) showing how the estimates decrease over time (scan number).

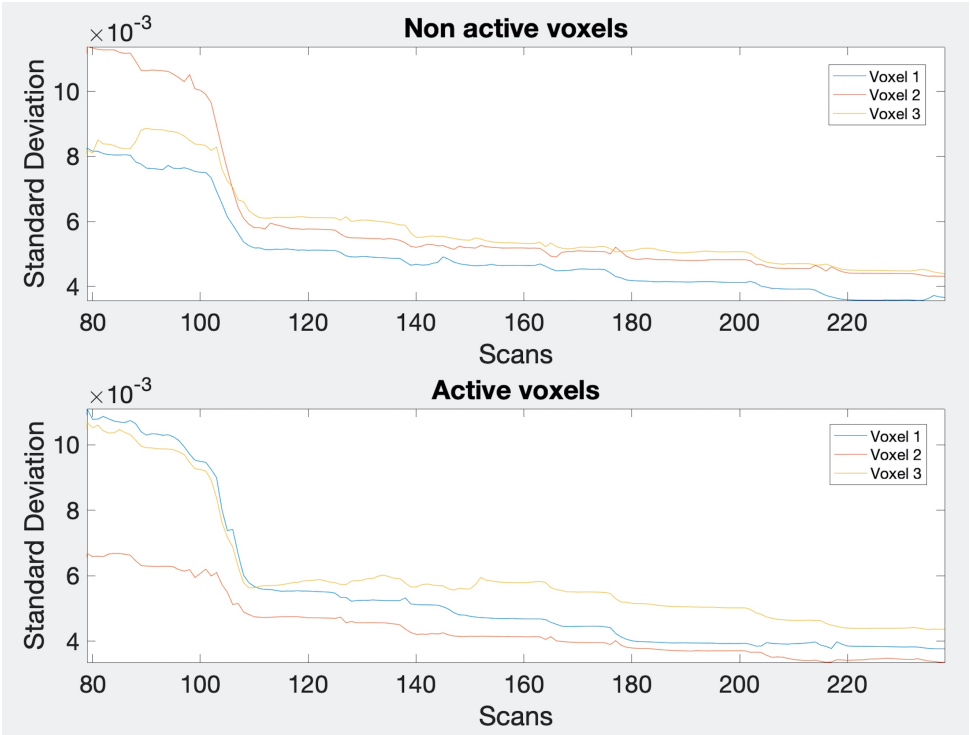

Figure 5: Plots showing the number of subjects that pass the framewise displacement threshold of 0.9 mm for each scan. Top: EPT subjects, bottom: control subjects.

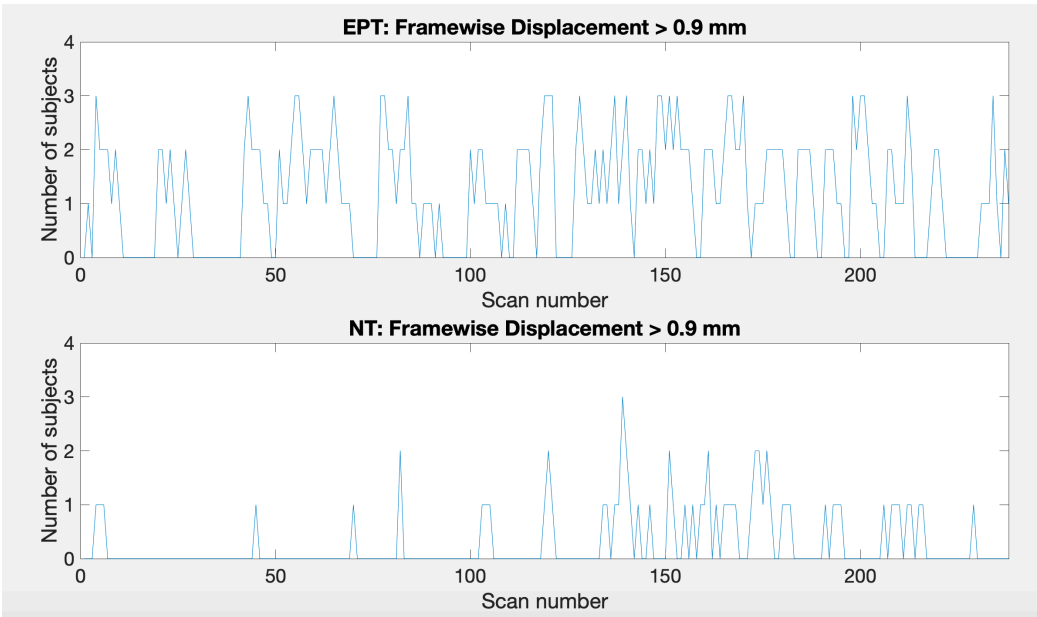

Figure 6: Group results for the 1-back task. Analysis performed for controls and EPT subjects using FSL. Early stopping with 2- and 4-blocks being initially administered is compared to full duration. Activations are overlaid on the MNI template brain. Red (A) = easy level results, Blue (B) hard level results. $P < 0.001$ uncorrected. Slices z = 58 is shown. R = right, L = left, A = anterior, P = posterior.

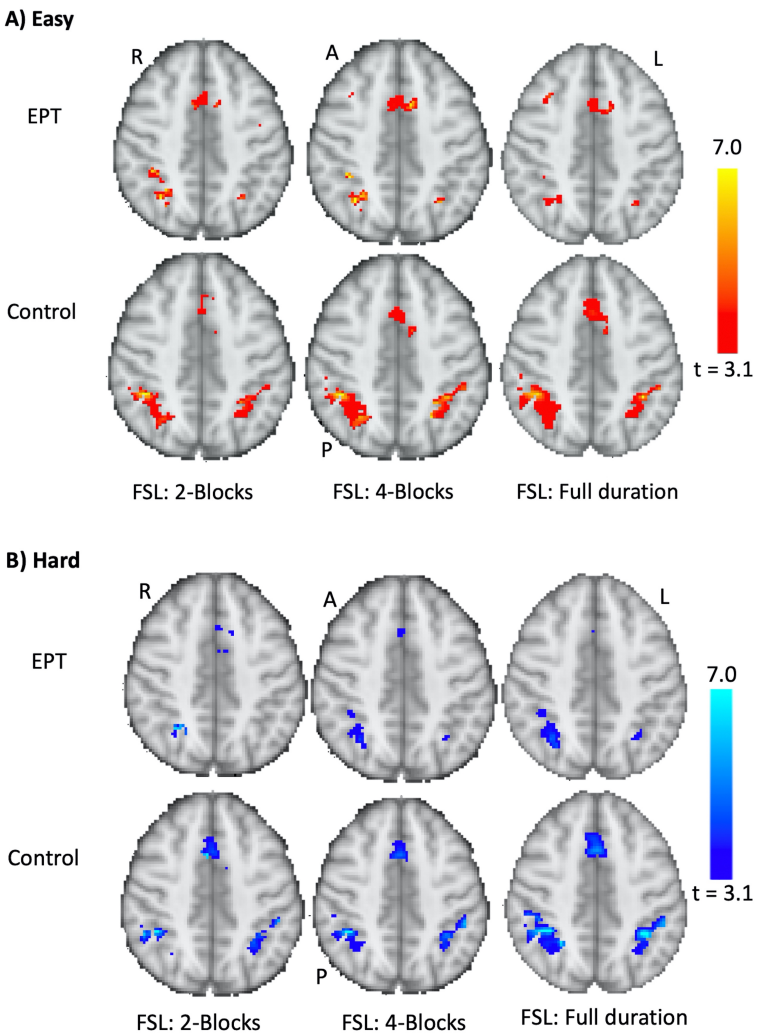

Table 1: The scan number when each stimulus block is completed.

Block	Easy level	Hard level
1	21	40
2	59	78
3	97	116
4	135	154
5	173	192
6	211	230

Table 2: Subject results of the 1-back task using SPRT to analyse the data. Analysis reported here uses $\alpha_E = 0.001$, $\beta_E = 0.1$ and thresholded at $p < 0.001$. A) Easy level: 2-block, B) Hard level: 2-block, C) Easy level: 4-block and D) Hard level: 4-block. At each subject's stopping scan, the number of active voxels that spatially overlap between early stopping and full duration are given. The percentage of voxels-in-common is also given relative to the total number of active voxels at the full duration scan. Maximum number of possible scans is 238, minimum is 78 scans for 2

blocks first stage of easy and hard stimulus administration or 154 scans for 4 blocks first stage of easy and hard. Median values are calculated with those who stopped early only.

A) Easy level – 2 block first stage

Subject	No of Voxels in ROI (full brain)	No of Voxels Classified Active at 80%	Scan when 80% Reached	No of Voxels Classified Active at Scan 238	Voxels in Stopping Scan but not in Scan 238	Voxels in Scan 238 but not Stopping Scan	% overlap
Control							
1	115,062	3,556	Not reached	3,556	0	0	100
2	113,564	12,803	Not reached	12,803	0	0	100
3	77,359	5,379	3E/3H	13,472	96	8,189	39.2
4	103,591	13,680	Not reached	13,680	0	0	100
5	114,260	7,482	5E/5H	10,540	170	3,228	69.4
6	121,353	979	2E/2H	4,674	547	4,242	9.2
7	107,406	1,394	3E/3H	4,115	246	2,967	27.9
8	106,267	7,068	5E/5H	7,124	1,278	1,334	81.3
9	121,195	349	2E/2H	9,329	257	9,237	1.0
10	96,565	7,340	Not reached	7,340	0	0	100
11	107,016	1,597	3E/2H	6,357	363	5,123	19.4
12	96,936	12,429	Not reached	12,429	0	0	100
		Median:	3E/2H				27.9 (SD 30.2)
EPT							
13	94,623	59	2E/2H	49	59	49	0.0
14	94,905	15,556	6E/6H	15,866	445	755	95.2
15	98,799	1,014	4E/4H	750	743	479	36.1
16	118,098	13,484	Not reached	13,484	0	0	100
17	124,749	4,723	2E/2H	3,487	3,531	2,295	34.2
18	97,437	717	2E/2H	3,925	373	3,581	8.8
19	135,379	7,402	Not reached	7,402	0	0	100
20	89,609	1,361	2E/2H	8,093	39	6,771	16.3
21	104,584	10,039	Not reached	10,039	0	0	100
22	114,201	13,817	Not reached	13,817	0	0	100
23	86,177	6,715	Not reached	6,715	0	0	100
		Median:	2E/2H				34.2 (SD 34.0)

B) Hard level – 2 blocks first stage

Subject	No of Voxels in ROI (full brain)	No of Voxels Classified Active at 80%	Scan when 80% Reached	No of Voxels Classified Active at Scan 238	Voxels in Stopping Scan but not in Scan 238	Voxels in Scan 238 but not Stopping Scan	% overlap
Control							
1	115,062	23,351	Not reached	23,351	0	0	100

2	113,564	20,749	Not reached	20,749	0	0	100
3	77,359	3,478	2E/2H	14,976	18	11,516	23.1
4	103,591	5,986	3E/2H	9,874	1,735	5,623	43.1
5	114,260	3,743	2E/2H	14,939	135	11,331	24.2
6	121,353	6,129	4E/3H	7,829	527	2,227	71.6
7	107,406	330	3E/2H	2,494	91	2,255	9.6
8	106,267	14,634	Not reached	14,634	0	0	100
9	121,195	3,253	3E/2H	8,898	412	6,057	31.9
10	96,565	4,948	3E/2H	10,921	2,114	8,087	26.0
11	107,016	11,355	Not reached	11,355	0	0	100
12	96,936	29,279	Not reached	29,279	0	0	100
		Median:	3E/2H				26.0 (SD 19.9)
EPT							
13	94,623	133	79	860	133	860	0.0
14	94,905	7,676	2E/2H	21,074	546	13,944	33.8
15	98,799	892	3E/2H	1,696	684	1,488	12.3
16	118,098	14,984	Not reached	14,984	0	0	100
17	124,749	1,796	2E/2H	4,335	918	3,457	20.3
18	97,437	10,967	Not reached	10,967	0	0	100
19	135,379	1,274	2E/2H	7,609	761	7,096	6.7
20	89,609	1,535	2E/2H	8,754	53	7,272	16.9
21	104,584	2,415	2E/2H	10,038	1,181	8,638	12.3
22	114,201	15,374	Not reached	15,374	0	0	100
23	86,177	8,088	Not reached	8,088	0	0	100
		Median:	2E/2H				14.6 (SD 9.4)

C) Easy level – 4 blocks first stage

Subject	**No of Voxels in ROI (full brain)**	**No of Voxels Classified Active at 80%**	**No of Easy/Hard blocks when 80% Reached**	**No of Voxels Classified Active at Scan 238**	**Voxels in Stopping Scan but not in Scan 238**	**Voxels in Scan 238 but not Stopping Scan**	**% overlap**
Control							
1	115,062	3,556	Not reached	3,556	0	0	100
2	113,564	12,803	Not reached	12,803	0	0	100
3	77,359	10,921	5E/4H	13,472	471	3,022	77.6
4	103,591	13,680	Not reached	13,680	0	0	100
5	114,260	7,482	5E/5H	10,540	170	3,228	69.4
6	121,353	3,806	5E/5H	4,674	603	1,471	68.5
7	107,406	2,195	5E/4H	4,115	99	2,019	50.9
8	106,267	7,068	5E/5H	7,124	1,278	1,334	81.3
9	121,195	4,065	4E/4H	9,329	127	5,391	42.2
10	96,565	7,340	Not reached	7,340	0	0	100
11	107,016	3,088	5E/5H	6,357	40	3,309	47.9

12	96,936	12,429	Not reached	12,429	0	0	100
		Median:	5E/4H (SD				68.5 (SD 15.4)
EPT							
13	94,623	139	4E/4H	49	136	46	95.2
14	94,905	15,556	6E/6H	15,866	445	755	33.5
15	98,799		4E/4H	750	648	499	100
16	118,098	13,484	Not reached	13,484	0	0	77.7
17	124,749	3,908	6E/6H	3,487	1,199	778	31.1
18	97,437	1,302	4E/4H	3,925	80	2,703	100
19	135,379	7,402	Not reached	7,402	0	0	14.3
20	89,609	1,302	4E/4H	8,093	148	2,898	100
21	104,584	10,039	Not reached	10,039	0	0	100
22	114,201	13,817	Not reached	13,817	0	0	100
23	86,177	6,715	Not reached	6,715	0	0	100
		Median:	4E/4H				33.5 (SD 34.3)

D) Hard level – 4 blocks first stage

Subject	**No of Voxels in ROI (full brain)**	**No of Voxels Classified Active at 80%**	**Scan when 80% Reached**	**No of Voxels Classified Active at Scan 238**	**Voxels in Stopping Scan but not in Scan 238**	**Voxels in Scan 238 but not Stopping Scan**	**% overlap**
Control							
1	115,062	23,351	Not reached	23,351	0	0	100
2	113,564	20,749	Not reached	20,749	0	0	100
3	77,359	13,160	6E/5H	14,976	67	1,883	87.4
4	103,591	9,874	Not reached	9,874	0	0	100
5	114,260	14,939	Not reached	14,939	0	0	100
6	121,353	6,129	5E/4H	7,829	527	2,227	71.6
7	107,406	1,214	4E/4H	2,494	202	1,482	40.6
8	106,267	14,634	Not reached	14,634	0	0	100
9	121,195	4,313	4E/4H	8,898	343	4,928	44.6
10	96,565	5,360	4E/4H	10,921	908	6,469	40.8
11	107,016	11,355	Not reached	11,355	0	0	100
12	96,936	29,279	Not reached	29,279	0	0	100
		Median:	4E/4H				44.6 (SD 21.3)
EPT							
13	94,623	1,919	4E/4H	860	1,532	473	45.0
14	94,905	20,638	4E/4H	21,074	2,774	3,210	84.8
15	98,799	1,696	Not reached	1,696	0	0	100
16	118,098	14,984	Not reached	14,984	0	0	100
17	124,749	4,482	6E/6H	4,335	1,101	954	78.0
18	97,437	10,967	Not reached	10,967	0	0	100
19	135,379	3,397	4E/4H	7,609	1,353	5,565	26.9

20	89,609	8,107	4E/4H	8,754	1,312	1,959	77.6
21	104,584	10,038	Not reached	10,038	0	0	100
22	114,201	15,374	Not reached	15,374	0	0	100
23	86,177	8,088	Not reached	8,088	0	0	100
		Median:	4E/4H				77.6 (SD 25.2)

Table 3: Overlap with full duration scan threshold of z = 4.0. The 2- and 4-block first stage results are thresholded at z = 3.1. The percentage of voxels-in-common is also given relative to the total number of active voxels at the full duration scan. Median values are calculated with those who stopped early only.

A) Easy level

Subject	**No of Voxels Classified Active at Scan 238**	**2-Blocks** **Voxels in Stopping Scan but not in Scan 238**	**Voxels in Scan 238 but not Stopping Scan**	**% overlap**	**4-Blocks** **Voxels in Stopping Scan but not in Scan 238**	**Voxels in Scan 238 but not Stopping Scan**	**% overlap**
Control							
1	1,099	0	0	100	0	0	100
2	8,166	0	0	100	0	0	100
3	10,818	229	5,668	47.6	1,400	1,297	88.0
4	8,535	0	0	100	0	0	100
5	6,730	1,178	426	93.7	1,178	426	93.7
6	2,598	610	2,229	14.2	1,527	319	87.7
7	2,313	420	1,339	42.1	451	569	75.4
8	4,035	3,232	199	95.1	3,232	199	95.1
9	5,836	297	5,784	0.9	401	2,172	62.8
10	4,077	0	0	100	0	0	100
11	3,487	721	2,611	25.1	343	742	78.7
12	7,413	0	0	100	0	0	100
		Median:	42.1 (SD 36.9)			87.7 (SD 11.5)	
EPT							
13	0	59	0	0	139	0	0
14	16,095	3,502	0	74.9	3,502	0	74.9
15	108	927	21	80.6	811	20	81.5
16	7,887	0	0	100	0	0	100
17	1,353	4,119	749	44.6	2,583	28	97.9
18	1,943	480	1,706	12.2	198	839	56.8
19	3,582	0	0	100	0	0	100
20	6,202	79	4,920	20.7	456	1,315	13.6
21	2,816	0	0	100	0	0	100
22	8,102	0	0	100	0	0	100
23	4,654	0	0	100	0	0	100

		Median:	32.7 (SD 33.5)			65.9 (SD 39.2)	

B) Hard level

Subject	No of Voxels Classified Active at Scan 238	2-Blocks Voxels in Stopping Scan but not in Scan 238	Voxels in Scan 238 but not Stopping Scan	% overlap	4-Blocks Voxels in Stopping Scan but not in Scan 238	Voxels in Scan 238 but not Stopping Scan	% overlap
Control							
1	15,071	0	0	100	0	0	100
2	13,253	0	0	100	0	0	100
3	12,197	68	8,787	28.0	1,101	138	98.9
4	6,566	2,613	3,193	51.4	0	0	100
5	10,367	314	6,938	33.1	0	0	100
6	5,131	1,609	611	88.1	1,609	611	88.1
7	1,193	146	1,009	15.4	367	346	71.0
8	9,802	0	0	100	0	0	100
9	5,984	590	3,321	44.5	673	2,344	60.8
10	6,147	2,828	4,027	34.5	1,686	2,473	59.8
11	6,331	0	0	100	0	0	100
12	20,299	0	0	100	0	0	100
		Median:	34.5 (SD 23.3)			71 (SD 17.2)	
EPT							
13	191	133	191	0	1,794	66	65.4
14	16,095	1,120	9,539	40.7	5,538	995	93.8
15	532	748	388	27.1	0	0	100
16	11,257	0	0	100	0	0	100
17	1,906	1,207	1,317	30.9	2,597	21	98.9
18	5,975	0	0	100	0	0	100
19	3,071	980	2,777	9.6	2,008	1,682	45.2
20	6,787	79	5,331	21.5	2,164	844	87.6
21	4,987	1,503	4,075	18.3	0	0	100
22	9,937	0	0	100	0	0	100
23	5,744	0	0	100	0	0	100
		Median:	21.5 (SD 13.6)			87.6 (22.4)	

Table 4: The number of active voxels that spatially overlap between early stopping and full duration group analyses are listed. Images thresholded at $p < 0.001$. The percentage of voxels-in-common is given relative to the total number of active voxels detected at full duration. Stopping based on 80% classification at the individual level.

	No of Active Voxels	Voxels in Stopping Scan but not in Scan 238	Voxels in Scan 238 but not Stopping Scan	% of Common Voxels with Scan 238	Variance Magnitudes
EPT - Easy					
2-Blocks	743	478	263	50.2	45,473 (SD 2,183)
4-Blocks	839	466	155	70.6	30,507 (SD 1,385)
Full duration	528				18,351 (SD 1,135)
EPT - Hard					
2-Blocks	633	377	930	21.6	33,115 (SD 2,209)
4-Blocks	1,002	225	155	65.5	28,230 (SD 1,170)
Full duration	1,186				27,110 (SD 1,054)
Controls - Easy					
2-Blocks	2,065	730	941	58.7	14,135 (SD 640)
4-Blocks	2,882	1,150	409	76.1	21,118 (SD 857)
Full duration	2,276				11,082 (SD 469)
Controls – Hard					
2-Blocks	2,691	1,535	1,364	45.9	20,756 (SD 1,008)
4-Blocks	2,433	789	876	65.2	11,594 (SD 743)
Full duration	2,520				9,104 (SD 685)