

Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses.

Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, Jean-Baptiste Poline

▶ **To cite this version:**

Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, et al.. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses.. NeuroImage, 2007, 35 (1), pp.105-20. 10.1016/j.neuroimage.2006.11.054 . cea-00371089

HAL Id: cea-00371089
https://cea.hal.science/cea-00371089

Submitted on 26 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analysis of fMRI data sampled from Large Populations: Statistical and Methodological Issues

Bertrand Thirion[†], Philippe Pinel[+], Sébastien Mériaux[*],
Alexis Roche[*], Stanislas Dehaene[+], Jean-Baptiste Poline[*]

May 31, 2006

[†]INRIA Futurs
Service Hospitalier Frédéric Joliot,
4, Place du Général Leclerc
91401 Orsay Cedex
E-mail: thirion@shfj.cea.fr

[*] Département de Recherche Médicale - CEA - DSV
Service Hospitalier Frédéric Joliot,
4, Place du Général Leclerc
91401 Orsay Cedex

[+] Unité INSERM 562 "Neuroimagerie Cognitive"
Service Hospitalier Frédéric Joliot,
4 Place du Général Leclerc
91401 Orsay cedex, France

Abstract

Validating the association between brain activity, as measured in functional MRI, with a combination or a contrast of tasks is usually performed by replicating an experiment in a small group of subjects, and by assessing the presence of a statistically significant average effect across subjects (random effects analyses). While many efforts have been made to control the rate of false detections, statistical characteristics of the data have rarely been studied, and the reliability of the results (supra-thresholds areas that are considered as activated regions) has rarely been assessed. In this work, we take advantage of the large cohort of subjects who underwent the *Localizer* experiment to study the statistical nature of group data, propose some measures of the reliability of group studies, and address simple methodological questions as : is there, from the point of view of reliability, an optimal statistical threshold for activity maps ? How many subjects should be included in group studies ? What method should be preferred for inference ? Our results suggest that *i)* optimal thresholds can indeed be found, and are rather lower than usual corrected for multiple comparison thresholds *ii)* 20 subjects or more should be included in functional neuroimaging studies in order to have sufficient reliability, *iii)* non-parametric significance assessment should be preferred to parametric methods *iv)* cluster-level thresholding is more reliable than voxel-based thresholding *v)* mixed effects tests are much more reliable than random effects tests. Moreover, our study shows that inter-subject variability plays a prominent role in the relatively low sensitivity and reliability of group studies.

1 Introduction

1.1 Inter-subject variability in neuroimaging

Many scientists spend lots of efforts in order to obtain statistically significant results in neuroimaging studies, in order to validate a prior hypothesis on brain function, and it is certainly true that one of the greatest difficulties that they have to face is the high variability that is present in their datasets across subjects. In particular, this high degree of variability dramatically reduces the sensitivity of random effects studies. For instance, in [Wei et al., 2004], it has been shown empirically that inter-session variability was much higher than inter-subject variability.

Several techniques have been employed to quantify inter-subject differences [Kherif et al., 2004]. The use of such tools shows that inhomogeneous populations can be met quite frequently in neuroimaging studies. This problem is dramatic in general, given the small number of subjects in many studies (10-15 subjects, sometimes less).

Analysing the variability that arises between different datasets is not straightforward, since it is probably of a mixture of random and deterministic or structured factors. The first one is inter-session variability, which is caused by different amount of motion, attention or tiredness of the subjects. However, several specific factors can be proposed in the inter-subject setting :

- Spatial mismatch between datasets. It is known that perfect correspondences between two anatomical images cannot be found in general, and that correspondences should generally be considered as approximative, even after rigid or non-rigid spatial normalization. The magnitude order of such local shifts is probably as large as 1cm in many instances (this can be observed for functional regions like the the motor cortex or the visual areas [Thirion et al., ress, Stiers et al., 2006] or the position of anatomical landmarks

[Collins et al., 1998, Hellier et al., 2003]). Across subjects, this effect typically yields a structured but variable pattern.

- Next, the magnitude of the BOLD signal recorded at the same location for several tasks is variable across subjects, and sometimes across sessions [Smith et al., 2005], and the precise nature of this variability is not clear. It should be reminded that fMRI is not a quantitative neuroimaging modality, and that the standard use of reporting percent of signal increase is also problematic, due to the ambiguous definition (voxel-based or global average) of the baseline reference.

- Finally, there could be global differences in the cascade of neural activation elicited by a given task, related to genetic or epigenetic differences across subjects, or to different cognitive strategies (for non trivial tasks). This is of course interesting in itself, but at the same time very difficult to account for in traditional studies, and in many instances very difficult to prove.

This is a very sketchy account of possible sources of variability across subjects, but one should remember that *all* these effects equally treated as *confounds* and globally modeled as second-level variability [Friston et al., 2002] terms in current linear models.

It follows that voxel-based random effects analyses, that assess the significance of an effect by comparing its mean value to its variability across subjects are typically very weakly sensitive [Friston et al., 1999, McNamee and Lazar, 2004]. Moreover the lack of reliability makes brain mapping procedures [Jernigan et al., 2003] problematic. We illustrate this issue in Fig. 1 with an example taken from the dataset presented next.

1.2 Methodology in group analyses

Voxel-based random effects analysis is the standard way to analyse data from group studies (although the extraction of discrete local maxima [Worsley, 2005] presents an attractive alternative). It is in fact a particular instance of mixed effects models, which recently gained in popularity [Worsley et al., 2002, Friston et al., 2002, Beckmann et al., 2003, Neumann and Lohmann, 2003, Woolrich et al., 2004]. However, there remain some alternatives, i.e. non-systematic parts, in the group data processing strategy.

- Choice of the threshold. Usually, statistical parametric maps (SPMs) are thresholded in such a way the the number or the rate of false positive voxels is controlled, at least asymptotically. In particular, the chosen threshold does not reflect a trade-off between the necessity of controlling both the number of false positive and the number of false negative. One straightforward reason is that it is relatively simple to model the statistical distribution under the null hypothesis, but not under the converse hypothesis.

- It is possible to threshold SPMs at the voxel or at the cluster level. In the latter case, a double thresholding procedure is used : first the map is thresholded at a (relatively lenient) significance level ; second, the size of the resulting connected components is assessed against its distribution under the null hypothesis. This is usually considered as a safe procedure, but fully neglects the possibility of small yet significant activation foci. One of the reasons is that intensive smoothing of the data simply removes such eventuality.

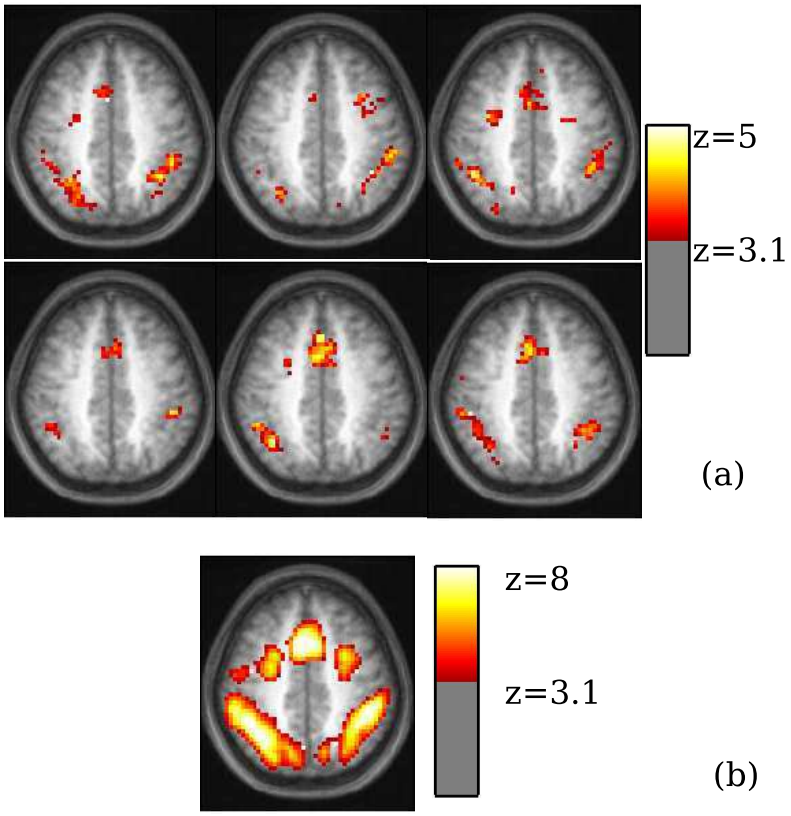


Figure 1: Illustration of the low sensitivity and weak reliability of supra-threshold patterns in standard group studies. (a) For a functional contrast that shows regions involved in a computation task, we show activity maps thresholded at a $p < 0.001$ level after a random effect analysis on 6 disjoint groups of 13 subjects; the position of the view is z=37mm in the MNI normalized space.(b) In the same plane, here is the same map computed from all the subjects together. Note the low sensitivity and weak reliability of the maps in (a).

- Another alternative consists in the choice of parametric versus non-parametric tests. While parametric tests are particularly efficient and computationally cheap, they are based on possibly unrealistic hypotheses that may reduce their sensitivity. For instance, assessing a Student statistic obtained in a group analysis with respect the Student distribution with the appropriate number of degrees of freedom assumes that the data is normally distributed under the null hypothesis. This hypothesis cannot be checked in the usual, small datasets. Non-parametric tests may avoid these issues [Holmes et al., 1996, Brammer et al., 1997, Bullmore et al., 1999, Nichols and Holmes, 2002, Hayasaka and Nichols, 2003, Mériaux et al., 2006a], but at a higher computational cost.

- Finally, many other alternatives concern the different pre-processing steps that are standard in typical studies. For simplicity, let us only retain the choice of a smoothing kernel (typically 8 to 12 mm FWHM). It is has been shown [Shaw et al., 2003, LaConte et al., 2003] that cross-validation schemes could help to optimize these pre-processing choices.

Some alternatives may be proposed to assess the significance of activity in group studies. For instance, parcellation, and thus parcel-based random effects maps can be used [Thirion et al., ress], with a possibly double advantage : if parcels adapt to individual anatomy, they can cope with some parts of the inter-subject variability ; second, this considerably alleviates the multiple comparison problem. Care must however be taken when controlling the test specificity. In general, making statistical inference at the voxel level is a particular, and not necessary optimal, choice.

However, another fundamental question has not received as much attention so far : it concerns the number of subjects that should be included in the study. This number typically represents a trade-off between (a) the cost of making neuroimaging experiments on large cohort of subjects and (b) the necessity to have enough subject for the significance of statistical tests [Desmond and Glover, 2002, Murphy and Garavan, 2004]. However, the true question could or should me : how many subjects are enough to make the analysis reliable, in terms of avoiding false positives as well as false negatives ? This questions has been raised in several studies, but reliability analysis has to our knowledge, not been used in this context.

Reliability can be thought of as the confidence level that can be given to a certain result. Following [Liou et al., 2003] we define it practically as the agreement between independent measurements of a certain effect.

1.3 Taking advantage of a large dataset

In this work, we use the large dataset collected in a localizer experiment (see the companion paper) to address empirically some of the questions listed above. First, we perform a relatively elementary statistical description of the dataset. In particular we focus on the deviation from normality in the distribution of effects for a given contrast, and on the spatial distribution of second-level variance in the dataset.

Then we concentrate on the reliability of the results when varying different features in the statistical procedure. Reliability analysis is based on binary i.e. thresholded maps obtained from distinct subgroups of subjects. One measurement consists in modelling the activated/inactivated status of the voxels in the image as a mixture of binary distributions, and to assess the reproducibility of the activated/inactivated status using this model. The second measure is based on

the distance between clusters of a certain size in the thresholded binary maps. A large distance means that no correspondence can be found between supra-threshold clusters across groups of subjects, hence that the maps are not very reliable. We use this framework to study the effect of the threshold value, then of the size of the group of subjects, and finally the effect of the method itself on the reliability of the maps. All this is possible due to the large sample size of the population of subjects. Importantly also, we show that our cross-validation procedure is not biased by possible confounds like a correlation between reliability and sensitivity, and that both reliability measures yield similar results.

An important point is that this procedure can be performed for different cognitive contrasts which have different characteristics in terms of contrast-to-noise ratio or spatial variability.

2 Materials and Methods

2.1 Dataset

The experimental paradigm is described with more detail in the companion paper. Briefly, we used an event-related experimental paradigm consisting of ten conditions. Subjects underwent a series of stimuli or were engaged in task such as passive viewing of horizontal or a vertical checkerboards, left or right click after audio or video instruction, computation (subtraction) after video or audio instruction, sentence listening, from audio or visual modality. Events were randomly occurring in time (mean inter stimulus interval: 3s), with ten occurrences per event type (except motor button clicks for which there are only five trials per session).

Eighty-one right-handed subjects participated in the study. The subjects gave informed consent and the protocol was approved by local ethics committee. Functional images were acquired on a 3T Bruker scanner using an EPI sequence (TR = $2400ms$, TE = $60ms$, matrix size=64×64, FOV = $24cm \times 24cm$). Each volume consisted of n_a $4mm$-thick axial slices without gap, where n_a varied from 26 to 40 according to the session. A session comprised 130 scans. The first four functional scans were discarded to allow the MR signal to reach steady state. Anatomical T1 images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2\ mm^3$.

fMRI data processing consisted in 1) temporal Fourier interpolation to correct for between-slice timing, 2) motion estimation. For all subjects, motion estimates were smaller than 1mm and 1 degree, 3) spatial normalization of the functional images, re-interpolation to $3 \times 3 \times 3\ mm^3$, and 4) smoothing (5mm FWHM). This pre-processing was performed with the SPM2 software (see e.g. [Ashburner et al., 2004]). Datasets were also analyzed using the SPM2 software, using standard high-pass filtering and AR(1) whitening. For further analysis, the voxel-based estimated effects for several contrasts of interest were retained.

We determined a global brain mask for the group by considering all the voxels that belong to at least half of the individual brain masks defined with SPM2. It comprises approximately 60000 voxels. Note that considering merely the intersection of the individual masks yields about 30000 voxels only.

2.2 Elementary Statistical description of the dataset

In this section, we select a few contrasts of interest, and study the statistical distribution of the corresponding parameters in each voxel. Using a first level (subject-specific) GLM, one can obtain parametric estimates of the BOLD activity at each voxel in each subject: For each subject $s \in [1..S]$ and each voxel $v \in [1..V]$, we have a parameter estimate $\hat{\beta}(s,v)$, and a variance estimate $\hat{\sigma}(s,v)^2$.

The first question that may arise is whether the effects $\hat{\beta}(s,v)$ are normally distributed or not, since this is a key assumption in classical (Random Effects) group studies. We have used the test of D'Agostino-Pearson [Zar, 1999], based on the computation of the skew and the kurtosis (third and forth order cumulants) of the values $\{\hat{\beta}(s,v)\}, s = 1..S$ in each voxel v. This provides us with a P-value of the D'Agostino-Pearson statistic under the null (normal) hypothesis. For the sake of visualization, we convert the P-value into a z-value. We have then repeated the procedure based on the normalized effects $\{\tau(s,v) = \frac{\hat{\beta}(s,v)}{\hat{\sigma}(s,v)}\}, s \in [1..S], v \in [1..V]$ which removes a potential variability in signal scaling across the population. Note that at the group level, the normalization through the residual magnitude has a much greater impact than the deviation from normality due to the fact that $\hat{\sigma}(s,v)$ is estimated with a finite ($\nu = 100$) number of degrees of freedom.

Then, assuming a two-level normal model of the data

$$\begin{aligned} \hat{\beta}(s,v) &= \beta(s,v) + \varepsilon(s,v), \text{ with } \varepsilon(s,v) \sim \mathcal{N}(0, \sigma(s,v)^2) & (1) \\ \beta(s,v) &= \bar{\beta}(v) + \zeta(s,v), \text{ with } \eta(s,v) \sim \mathcal{N}(0, v_g(v)^2) & (2) \end{aligned}$$

where $\beta(s,v)$ is the true effect for subject s, $\hat{\beta}(s,v)$ is the estimated effect for subject s, and $\bar{\beta}(v)$ is the average effect in the population at voxel v; $\varepsilon(s,v)$ and $\eta(s,v)$ are first-level (estimation) and second-level (inter-subject) normal residual terms. The first equation represents thus the subject-specific estimation of the signal and the second one the group-level model. We have estimated the second level variance v_g in each voxel, since it plays a central role in many group-level statistics. In particular, an interesting question is whether $\bar{\beta}(v)$ and $v_g(v)$ are independent or not. Note that v_g is estimated by maximizing the likelihood of the data given v_g. Newton or EM estimation schemes can be used [Worsley et al., 2002, Mériaux et al., 2006b]. In this work, we use a Newton estimation scheme.

2.3 Group-level analysis methods: Voxel-based statistics

We review here different techniques used for voxel-based inter-subject activation detection. We consider a given contrast of interest. For each subject $s \in [1..S]$ and each voxel $v \in [1..V]$, we have a parameter estimate $\hat{\beta}(s,v)$, and a variance estimate $\hat{\sigma}(s,v)^2$.

A random Effects (RFX) statistic is based on model (1)-(2), in which the first level variance is neglected. It is defined as

$$\rho(v) = \frac{\text{mean}_{s\in[1..S]}\hat{\beta}(s,v)}{\sqrt{(S-1)\text{var}_{s\in[1..S]}\hat{\beta}(s,v)}} \quad (3)$$

Under the null hypothesis, assuming a normal distribution for $\{\hat{\beta}(s,v)\}, s = 1..S$, $\rho(v)$ is Student-distributed with $(S-1)$ degrees of freedom, and the P-value under the null hypothesis can be assessed with or without correction for multiple comparisons[1]. Alternatively, a non-parametric scheme can be used to estimate the distribution of $\rho(v)$ under the null hypotheses, based on milder assumptions [Hayasaka and Nichols, 2003, Mériaux et al., 2006a]. In this work, we use the analytical threshold.

A Mixed Effects (MFX) statistic takes into account the first-level variance: assuming a group or (second-level) variance $v_g(v)$ at each voxel v, the MFX statistic writes:

$$\mu(v) = \sum_{s=1}^{S} \frac{\hat{\beta}(s,v)}{\hat{\sigma}(s,v)^2 + v_g(v)} \left(\sum_{s=1}^{S} \frac{1}{\hat{\sigma}(s,v)^2 + v_g(v)} \right)^{-\frac{1}{2}} \tag{4}$$

Intuitively, MFX may perform better than RFX since it down-weights the observations with high first-level variance. The distribution of the quantity $\mu(v)$ under the null hypothesis is difficult to assess [Woolrich et al., 2004]. We rely on an non-parametric scheme as in [Mériaux et al., 2006a, Mériaux et al., 2006b]: we tabulate the values of $\mu(v)$ for different sign swaps of each subject's dataset in order to generate a distribution under the null hypothesis, and compare the actual values with their null distribution. A quicker but very conservative approximation ($\mu \sim t_{S-1}$, t_{S-1} being the Student law with $S-1$ degrees of freedom) is also possible.

A possible alternative consists in neglecting the group variance v_g in the previous formula. This yields a pseudo-MFX statistic, which we denote henceforth as ΨFX:

$$\Psi(v) = \sum_{s=1}^{S} \frac{\hat{\beta}(s,v)}{\hat{\sigma}(s,v)^2} \left(\sum_{s=1}^{S} \frac{1}{\hat{\sigma}(s,v)^2} \right)^{-\frac{1}{2}} \tag{5}$$

Note that this is the statistic proposed in [Neumann and Lohmann, 2003]. The difference is that we perform a frequentist test by estimating the distribution of $\Psi(v)$ under the null hypothesis by random sign swaps (which we refer to as non-parametric approach), exactly as for the MFX test.

As an alternative, we also use Wilkoxon's signed rank statistic (WKX) [Hollander and Wolfe, 1999], which starts with sorting the absolute effects in ascending order, then sums up the ranks modulated by the corresponding effect's sign, yielding:

$$W(v) = \sum_{s=1}^{S} sign(\hat{\beta}(s,v)) rank(\hat{\beta}(s,v)) \tag{6}$$

This statistic is data-independent, thus its significance is assessed very easily. Unlike the previous statistics, it does not assess the positivity of the average effect, but the asymmetry of the estimated effects $\hat{\beta}(s,v)$ with respect to 0, the null hypothesis being that $\hat{\beta}(s,v)$ are distributed symmetrically about 0. The main interest of this statistic is that it is not based on the hypothesis that the $(\hat{\beta}(s,v)), s = 1..S$ are normally distributed.

[1]Given our definition of the group mask, it may occur that functional data is available in a subsample of the population of size n, with $\frac{S}{2} \leq n \leq S$. In such a case, S is replaced by n in the formulas. In the present work, we systematically apply such corrections, even though we do not mention them.

2.4 Group-level analysis methods: Higher-level statistics

Higher-level or non voxel-based analyses statistical inference methods include cluster-based inference and parcel-based inference.

Cluster-based inference [Hayasaka and Nichols, 2003, Mériaux et al., 2006a] is simply an extension of the voxel-based procedures, based on a double thresholding of a statistic map: first, a threshold is performed at the voxel level, then supra-threshold clusters are kept whenever their size is statistically large enough. In our implementation, we measure connectivity using the 18-nearest neighbours of each voxel in 3D, and estimate the P-values at the cluster-level using the non-parametric framework.

Parcel-based inference is a different scheme in which parcels are defined across subjects using anatomical and/or functional information. Two possible schemes have been presented in [Flandin et al., 2002] and [Thirion et al., ress], based on a Gaussian Mixture Model (GMM) and a hierarchical approach respectively. A key issue of both techniques is to obtain functionally, but also spatially connected parcels. Statistics, e.g. RFX, can then be computed at the parcel level by working on parcel-based signal average instead of voxel-based signal (PRFX statistic). The advantage is that some spatial relaxation is possible in the definition of the parcels, allowing for a better spatial registration of functional information. A related difficulty arises when the same functional data is used to build the parcels and perform the test, since false positives may arise from the spatial relaxation procedure. The present work therefore aims at checking the reliability of PRFX maps.

2.5 Assessing the reliability of activation maps

In this work, we propose two measures to assess the reliability of the activation maps derived from group analysis. This first one, based on mixture of binomial distribution, characterizes the stability of the status (active/inactive) of each voxel of the dataset. The second one measures how frequently clusters of voxels are found at similar locations in the normalized MNI/Talairach space across subjects. We use these measures in a bootstrap framework that enable us to characterize the reproducibility of activation maps obtained at the group level.

2.5.1 Reliability measure at the voxel level

In order to estimate the reliability of a statistical model, we need a method to compare statistical maps issued from the same technique, but sampled from different groups of subjects. We use the reliability indexes elaborated in [Genovese et al., 1997, Liou et al., 2003, Liou et al., 2005]. Assume that a statistical procedure (e.g. thresholding) yields binary maps $g_1, .., g_R$ for different groups of subjects. At each voxel v, $[g_1(v), .., g_R(v)]$ is thus a R-dimensional binary vector. At the image level, the distribution of $G(v) = \sum_{r=1}^{R} g_r(v)$ is modelled by a mixture of two binary distributions, one for the null hypothesis, one for the converse hypothesis: Let π_A^1 be the probability that a truly active voxel is declared active, $\pi_A^0 = 1 - \pi_A^0$ the probability that a truly active voxel is declared inactive, π_I^1, the probability that a truly inactive voxel is declared active, $\pi_I^0 = 1 - \pi_I^1$ the probability that an truly inactive voxel is declared inactive, and λ the proportion of truly activated voxels. Then, using a spatial independence assumption, the log-likelihood of the data

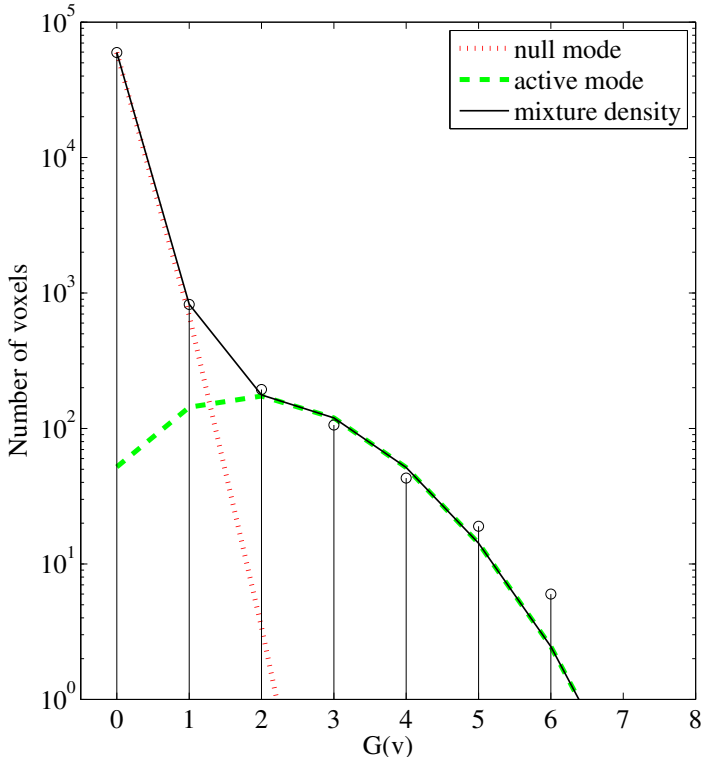


Figure 2: Example of mixture of binomial distribution. The empirical histogram of $G(v)$ is modelled by the model in Eq. (7), with $R = 8$. The Y axis is in log-coordinates for the sake of readability.

writes

$$\log(P(G)|\lambda, \pi_A^0, \pi_I^0) = cst + \sum_{v=1}^{V} \log\left(\lambda(\pi_A^0)^{R-G(v)}(\pi_A^1)^{G(v)} + (1-\lambda)(\pi_I^0)^{R-G(v)}(\pi_I^1)^{G(v)}\right) \quad (7)$$

Assuming $R \geq 3$ the three free parameters, $\pi_A^0, \pi_I^0, \lambda$ can be estimated using EM or Newton's methods. Note that optimizing the model over its different parameters sequentially, and using an adequate initialization, we could run the model for $R = 2$, though with higher variability in the estimation. An example of mixture model is given in Fig 2.

Given these estimates, the coherence index κ, known as Cohen's kappa is computed to measure the concordance of the different observations. Let $p_0 = \lambda\pi_A^1 + (1-\lambda)\pi_I^0$ be the proportion of agreement between the observations and the model. p_0 should be compared to the proportion of agreement by chance $p_C = \lambda\pi^0 + (1-\lambda)(1-\pi^0)$, where $\pi^0 = \lambda\pi_A^0 + (1-\lambda)\pi_I^0$ is the proportion of voxels declared inactive. The proportion of agreement corrected for chance is thus

$$\kappa = \frac{p_0 - p_C}{1 - p_C} \quad (8)$$

In this setting, $0 \leq \kappa \leq 1$ measures the accordance of the bimodal model with the data, which in turns reflects the coherence of the binary maps given as input to the model (7). If κ is close to 0, there is a very little agreement on which voxels are active, while there is a very good agreement if κ is close to 1. λ can also be kept as an index of the test sensitivity.

Note that more complex -and realistic- models have been proposed in the literature [Maitra et al., 2002], in which the parameter λ is allowed to vary spatially . However, our main purpose is not activation detection, but to obtain a global reliability measurement; for this reason, we keep the basic setting.

2.5.2 Reliability measure at the cluster level

Another way to assess the reliability of the results is to compare the positions of the clusters of supra-threshold voxels that arise through any group analysis. Assuming that the binary maps $g_1, .., g_R$ are obtained from different groups of subjects through a thresholding procedure, one can post-process them in order to yield connected components. The connected components with a size greater than a given threshold η are then retained, and their centre of mass (cm) is computed: let $x_i^r, i = 1..I(r)$ be the spatial coordinates of the cms derived from map g_r, we propose the following average distance between any two maps:

$$\Phi = \frac{1}{R(R-1)} \sum_{r=1}^{R} \sum_{s \in (1..R)-\{r\}} \frac{1}{I(r)} \sum_{i=1}^{I(r)} min_{j \in (1..I(s))} \phi(\|x_i^r - x_j^s\|), \tag{9}$$

where $\phi(x) = 1 - \exp\left(-\frac{x^2}{2\delta^2}\right)$ is a penalty function that is close to zero when the cluster centroids are properly matched and close to 1 otherwise. Φ represents the average mismatch between the cm of a supra-threshold component in a given map and the closest cm of any supra-threshold cluster obtained from another map. Appropriate penalty terms are used to handle the case $I(r) = 0$. We have performed some experiments using $\eta = 10$ voxels or $\eta = 30$ voxels, and use $\delta = 6mm$.

2.5.3 Procedure for the assessment of reliability

The procedure consists in dividing the population of 81 subjects in $R = 2, 3, 4, 5, 6$ or 8 disjoint groups of $\mathcal{S} = 40, 27, 20, 16, 13$ and 10 subjects respectively. The computation of different statistics, the derivation of an adequate threshold and the thresholding are performed in the different subgroups, and global reliability measures are derived from the ensuing binary maps. This procedure is repeated 100 times for each instance, yielding a distribution of the indexes κ, λ and Φ for each possible technique/parameter.

First, we choose the traditional RFX analysis procedure and adequate parameters and evaluate the distribution of the different indexes for three contrasts of interest. This is important to understand how well the indexes are characteristic of the amount, the spread and the variability of supra-threshold activity. In particular, it is important that the estimated reliability indexes are e.g. less variable for a given contrast than across contrasts.

Second, we evaluate the choice of the threshold on the different indexes, in the case of the voxel-based RFX test. While the sensitivity index certainly decreases while the threshold increases, the

behaviour of the reliability may be more complex, due to the trade-off between false positive and false negatives rates.

Third, we study the behavior of the different measurements when the number of subjects in the group varies; while it is obvious that reliability increases with the group size, it is not clear whether there exists a plateau and at which level. Previous studies [Desmond and Glover, 2002, Murphy and Garavan, 2004] suggest a steady increase of sensitivity with the group size.

Finally, we choose the following statistics: RFX, RFX on smoothed (12mm FWHM instead of 5mm) effect maps (SRFX), MFX, Wilkoxon(WKX), Cluster-level RFX (CRFX), Parcel-based RFX (PRFX) and ΨFX. RFX, SRFX, MFX, ΨFX and PRFX maps are thresholded at the $p < 0.001$ level, uncorrected for multiple comparisons. CRFX is thresholded at $p < 0.01$, uncorrected level at the voxel level, then at $p < 0.01$, at the cluster level.

PRFX maps are computed for a number $Q = 500$ of parcels. Since the parcel centres are defined at the group level in Talairach space, the voxels in the group result map are assigned to the parcel with the closest center in Talairach space. This results in a piecewise constant map, the pieces resulting from a Voronoi parcellation of the group mask into parcels. Note that in our bootstrap procedure, such boundaries are defined independently in each subgroup of subject. For parcellation, we use the hierarchical procedure presented in [Thirion et al., ress]; from our experiments (not shown), it yields slightly better results than the GMM procedure [Flandin et al., 2002].

3 Results

3.1 Statistical model of the inter-subject data

We have performed the D'Agostino-Pearson test on the effects $\hat{\beta}(v)$ of all the voxels, as well as the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}(v)$. This yields a map for each contrast. We present three of them, for the contrasts *left click-right click*, *audio instructions-video instructions* and *computation-reading*, thresholded at the $p < 0.001$ uncorrected level. We also present the inter-subject variability maps $v_g(v)$ computed in the mixed-effect model (2). We present these maps together with the RFX map based on 81 subjects in Fig. 3-5.

In either case, the regions with highest group variance are found in the regions with highest RFX statistics in absolute values; some of them are absent in the maps 3-5, where signed statistics are presented.

The observation of these maps suggests that

- High variance areas tend to co-localize with the activated areas. This implies that the parameters $v_g(v)$ and $\bar{\beta}(v)$ are certainly not independent, and that statistics that are penalized by the group variance may not be very efficient in general.
- Non-normality is very significant in wide regions of the brain.
- Deviation from the normality hypothesis is much lower for the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}$ than for the raw effects $\hat{\beta}$. It means dimensionless first-level statistics yield more homogeneous quantities across subjects than effects expressed in percents of baseline signal increase.

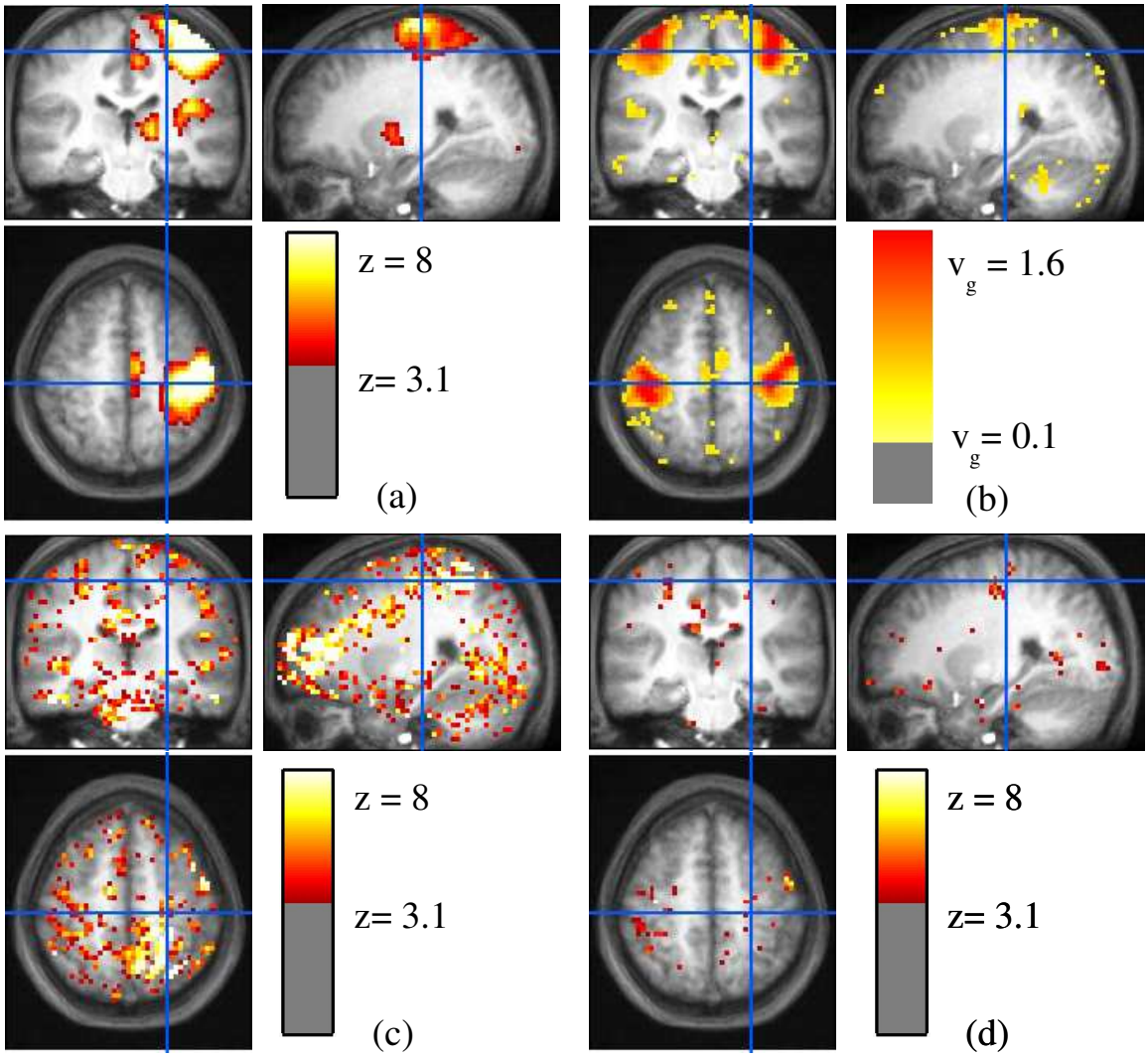


Figure 3: Statistical model of the effects for the *left click-right click* contrast, on $S = 81$ subjects. (a) z-value associated withe the RFX test; (b) group variance estimate; (c) z-value of the D'Agostino-Pearson test for normality of the effects $\hat{\beta}$; (d) z-value of the D'Agostino-Pearson test applied to the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the z-maps are thresholded at $z = 8$ for numerical reasons. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal.

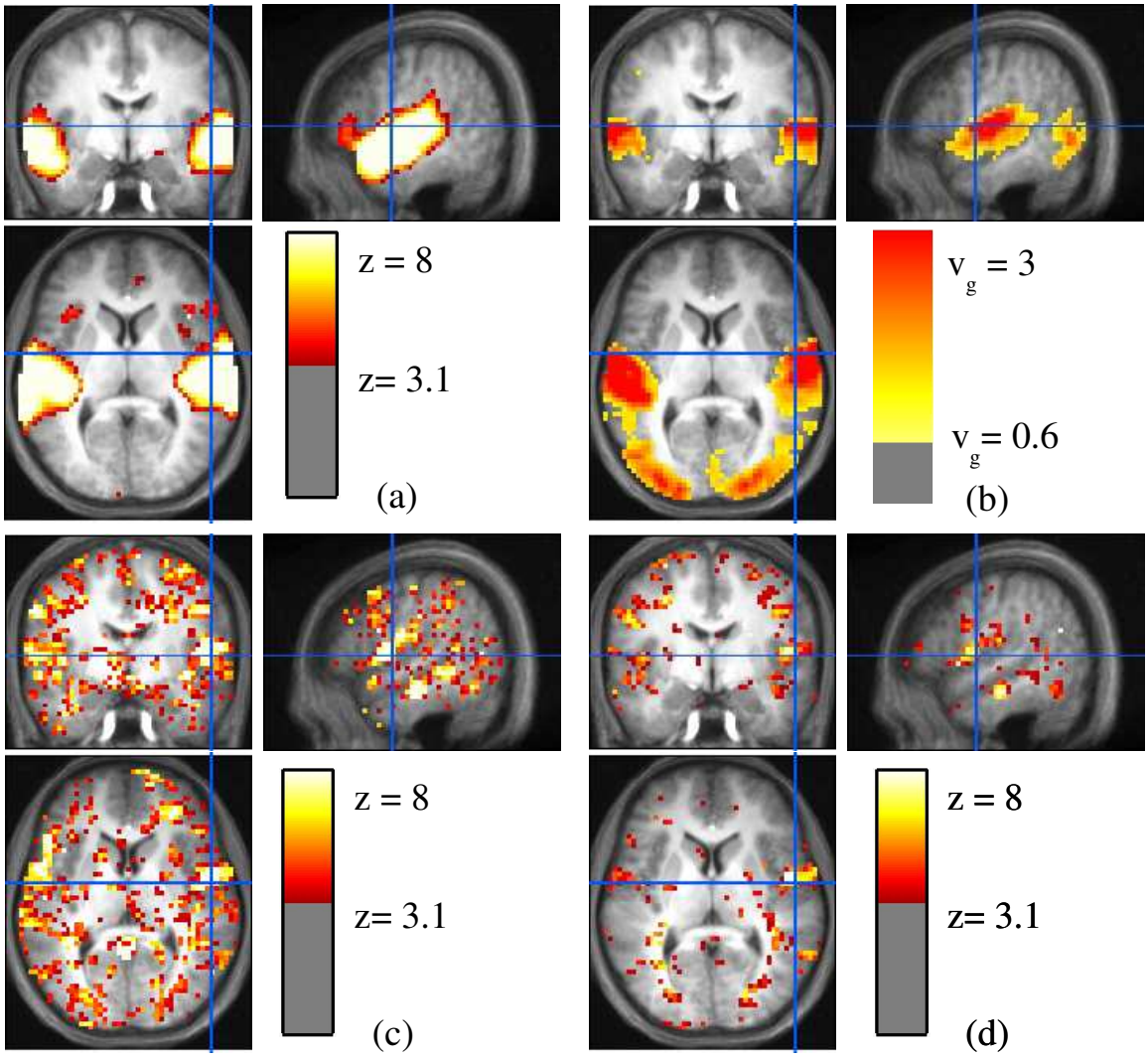


Figure 4: Statistical model of the effects for the *audio instructions-video instructions* contrast, on $S = 81$ subjects. (a) z-value associated withe the RFX test; (b) group variance estimate; (c) z-value of the D'Agostino-Pearson test for normality of the effects $\hat{\beta}$; (d) z-value of the D'Agostino-Pearson test applied to the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the z-maps are thresholded at $z = 8$ for numerical reasons. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal.

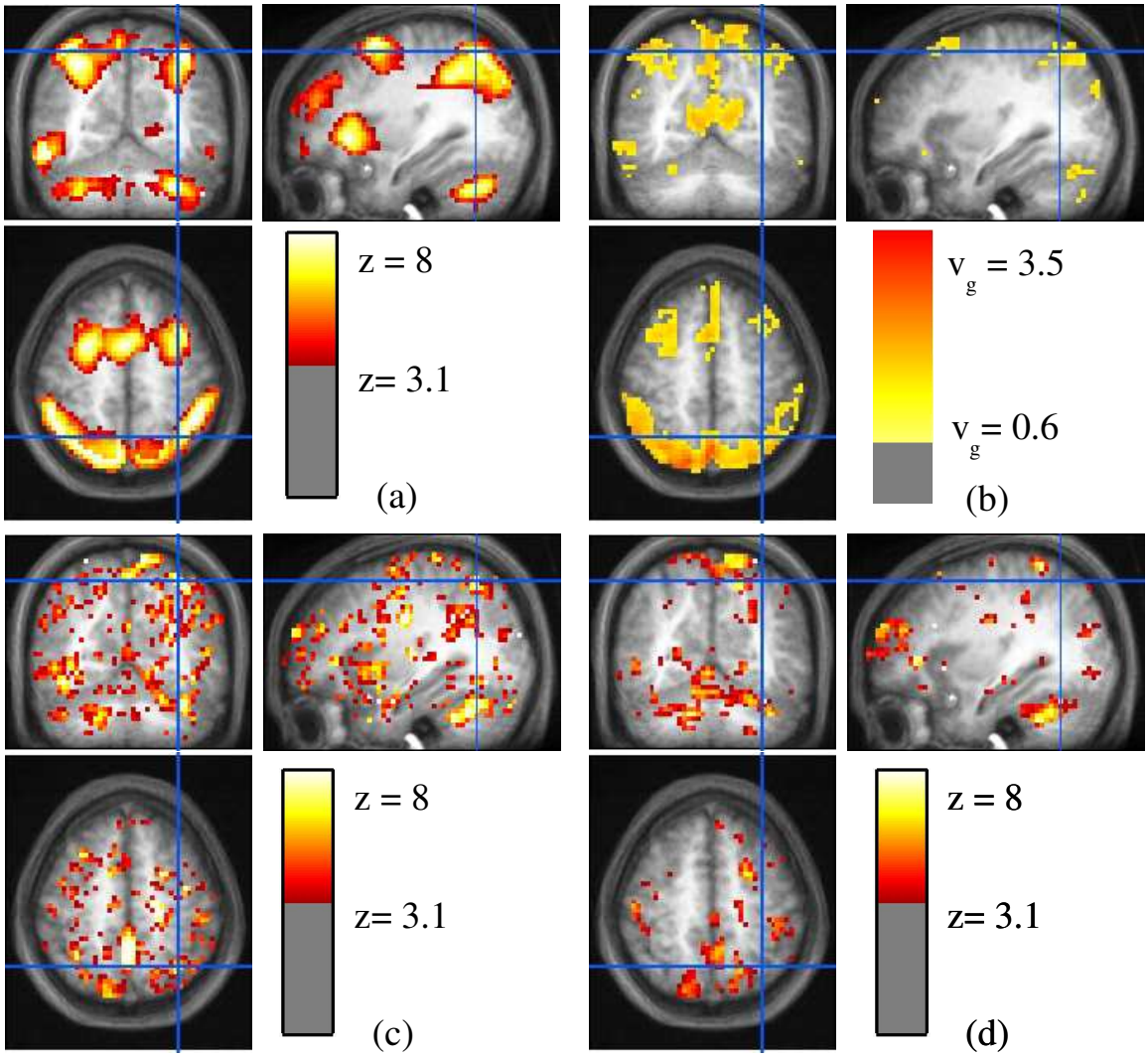


Figure 5: Statistical model of the effects for the *computation-reading* contrast, on $S = 81$ subjects. (a) z-value associated withe the RFX test; (b) group variance estimate; (c) z-value of the D'Agostino-Pearson test for normality of the effects $\hat{\beta}$; (d) z-value of the D'Agostino-Pearson test applied to the normalized effects $\frac{\hat{\beta}}{\hat{\sigma}}$. Note that all the z-maps are thresholded at $z = 8$ for numerical reasons. The color scale of the variance image has been chosen arbitrarily in order to have supra-threshold areas that are comparable with the other maps. The variance is expressed in squared percentage of the BOLD mean signal.

- Deviation from normality of the effects does not specifically co-localize with activated areas, but, in several cases it coincides with the boundaries of activated areas.

3.2 Reliability measurements for different cognitive contrasts

We have applied an RFX analysis for different cognitive contrasts, using $R = 5$ groups of $\mathcal{S} = 16$ and an RFX threshold $\theta = 3.1$. The contrasts are *left click-right click*, *audio instructions-video instructions* and *computation-reading*. The reliability index κ, the proportion of putative true positives λ and the inter-cluster distance penalty Φ are given in Fig. 6. It shows that κ and λ have a different behaviour and are strongly dependent on the cognitive contrast under study. For instance, the left motor contrast activates relatively small regions with a relatively low reliability; the auditory-selective contrast activates larger regions, with high reproducibility ; the computation-selective contrast activates larger regions, but with low reliability. The inter-cluster distance penalty Φ does not discriminate between the different contrasts as strongly as κ. As could have been expected, it has the opposite behaviour (maximal for the computation contrast, minimal for the auditory contrast).

3.3 How does the threshold affect the reliability of the analysis

Here we study the behaviour of our reliability measures when applied to a thresholded RFX map, when we let the threshold vary. The reliability measure is computed for 100 different splits of the population of subjects into $R = 5$ groups of $\mathcal{S} = 16$ subjects, in the case of the *left click-right click* contrast. The threshold varies from $\theta = 2.2$ ($p < 0.015$, uncorrected) to $\theta = 4$ ($p < 3.2.10^{-5}$,uncorrected) in steps of 0.2.

Logically, the sensitivity parameter λ decreases when θ increases (see Fig. 7(b)). More interestingly, κ reaches a maximum for $\theta^\star \sim 2.7$, but the index remains close at least for $\theta < 3.5$ as can be seen in seen in Fig. 7(a). Accordingly, the inter-cluster distance penalty Φ is minimized for a threshold $\theta^\star \sim 3$. The correspondence of this results is interesting, given that these two similarity measures are obtained independently, and based on different considerations. Note that we have obtained similar results when studying the other contrasts with slightly higher (auditory contrast) or lower (computation contrast) threshold values. Thereafter, we retain the threshold $\theta = 3.1$ ($p < 0.001$, uncorrected for multiple comparisons) for RFX analyses.

3.4 How many subjects are necessary to obtain a reliable RFX map

We study the dependence of κ, λ and Φ when we let the size $\mathcal{S}$ of the group vary. We base our investigation on the *left click-right click* contrast, with group maps thresholded at the $\theta = 3.1$ level. The results are presented in Fig. 8. It shows that the reliability increases with the group size, which was not unexpected. The sensitivity also increases with the group size. Interestingly, the reliability reaches a plateau only for $\mathcal{S} > 20$. The inter-cluster distance penalty Φ has a similar behaviour, with a plateau for $\mathcal{S} > 20$ subjects when $\eta = 10$, while lower values are reached when using $\eta = 30$.

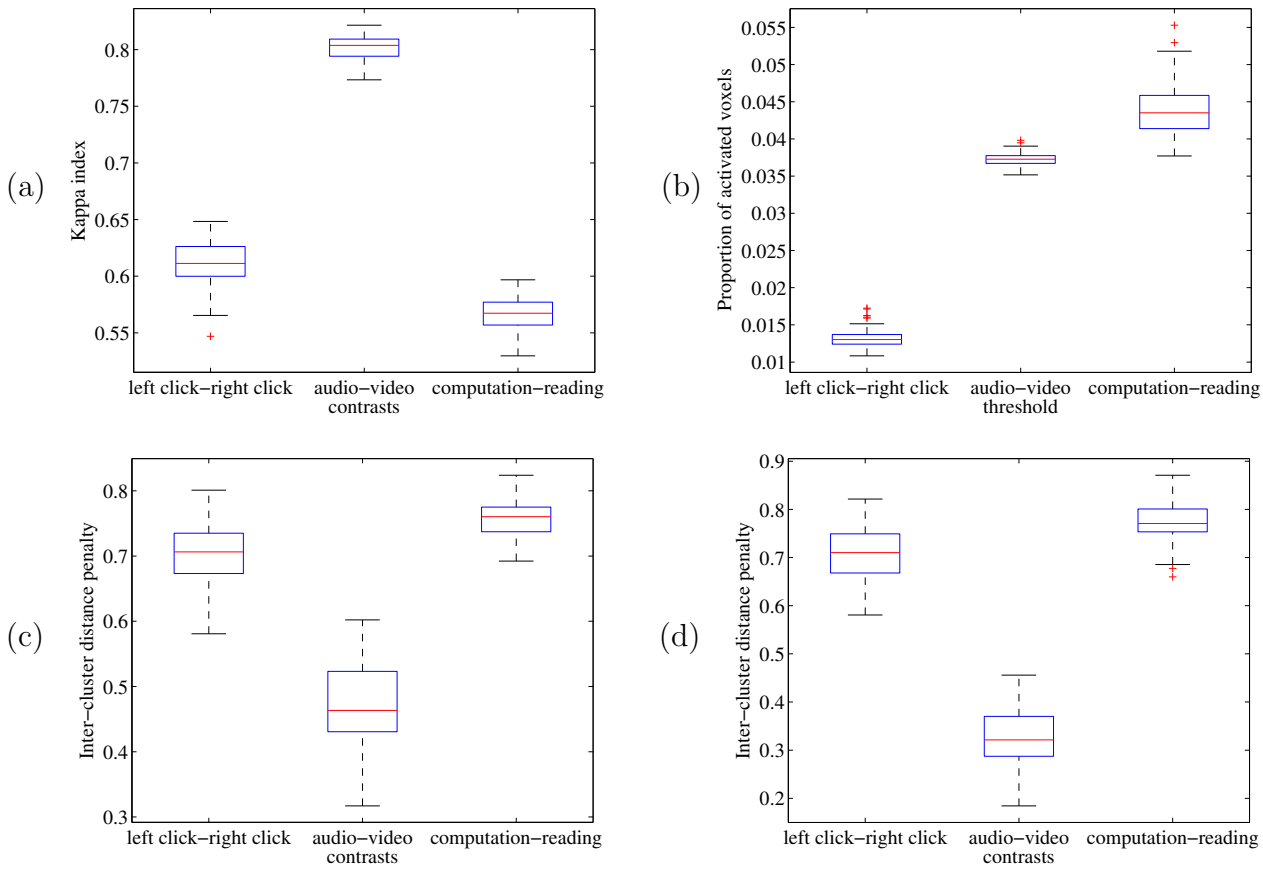


Figure 6: Dependence of the reproducibility and of the sensitivity of the group RFX analysis on the functional contrast under consideration. These results are obtained by drawing 5 disjoint groups of $\mathcal{S} = 16$ subjects in the population of 81 subjects, and applying the procedure described in section 2.5.1. The threshold is $\theta = 3.1$ (a) The results are more reliable for the contrast that shows auditory regions than for a contrast that shows motor activity or a contrast that shows the regions involved in the computation task. (b) By contrast the size of the putatively activated areas is greater for the contrast that shows regions involved in computation, and smaller for the contrast that shows the regions involved in motor activity. (c-d) The cluster variability penalty Φ is presented for clusters of more than $\eta = 10$ (c) or $\eta = 30$ (d) voxels. The behaviour is as expected, with the smallest value for the auditory-specific contrast.

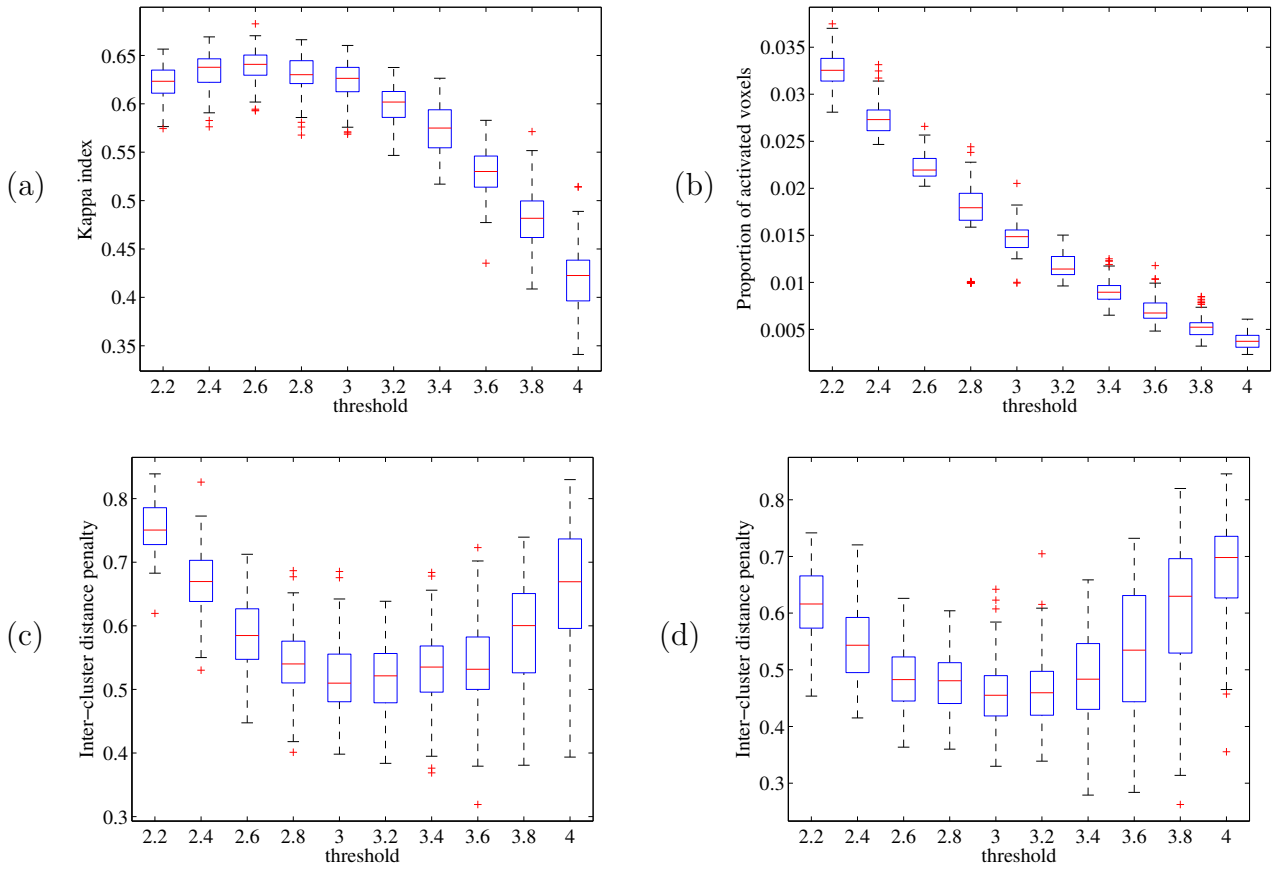


Figure 7: Dependence of the reproducibility, the sensitivity, and the distance between supra-threshold clusters of the group RFX analysis on the threshold chosen to binarize the statistic maps. These results are obtained by drawing 5 disjoint groups of $\mathcal{S} = 16$ subjects in the population of 81 subjects, and applying the procedure described in section 2.5.1. This is performed on the images of the *left click-right click* contrast. (a) The reproducibility index κ shows a maximum for $\theta \sim 2.7$. (b) The sensitivity decreases when θ increases. (c,d) The average distance between supra-threshold clusters of more than 10(c) or 30(d) voxels across groups has a minimum around $\theta \sim 3$.

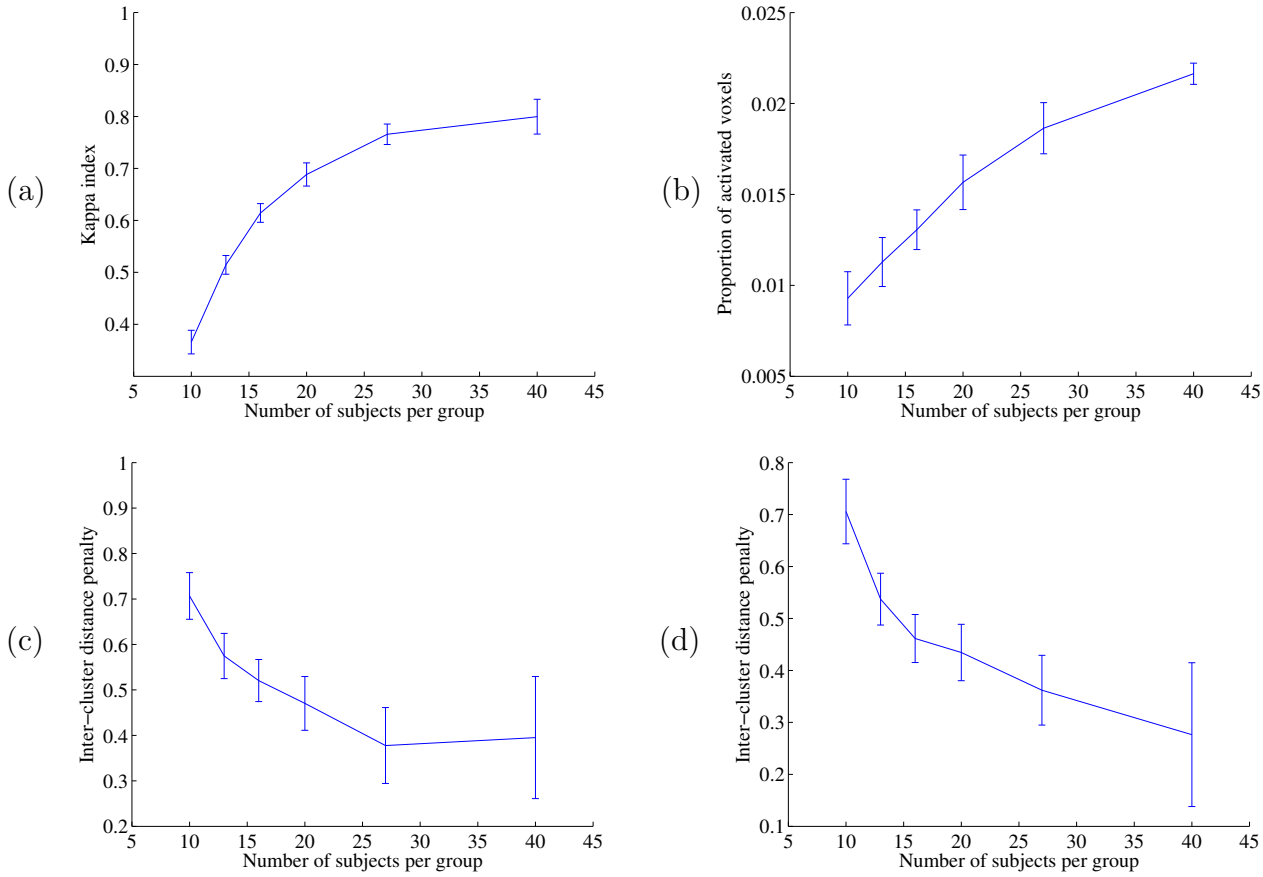


Figure 8: Dependence of the reproducibility κ (a), the sensitivity λ (b) and the average distance between supra-threshold cluster centroids Φ (c) of the group RFX analysis on the group size. The reliability is assessed considering disjoint groups of size $\mathcal{S} = 10, 13, 16, 20, 27, 40$ within the population of 81 subjects. This is performed on the images of the *left click-right click* contrast. (a) The reproducibility increases with $\mathcal{S}$ and reaches a plateau for $\mathcal{S} > 20$. (b) The size of putatively activated areas steadily increases with $\mathcal{S}$. (c-d) The average intra-cluster distance decreases with $\mathcal{S}$; it reaches a plateau for $\mathcal{S} > 20$ when $\eta = 10$ (c), whereas it further decreases when $\eta = 30$ (d).

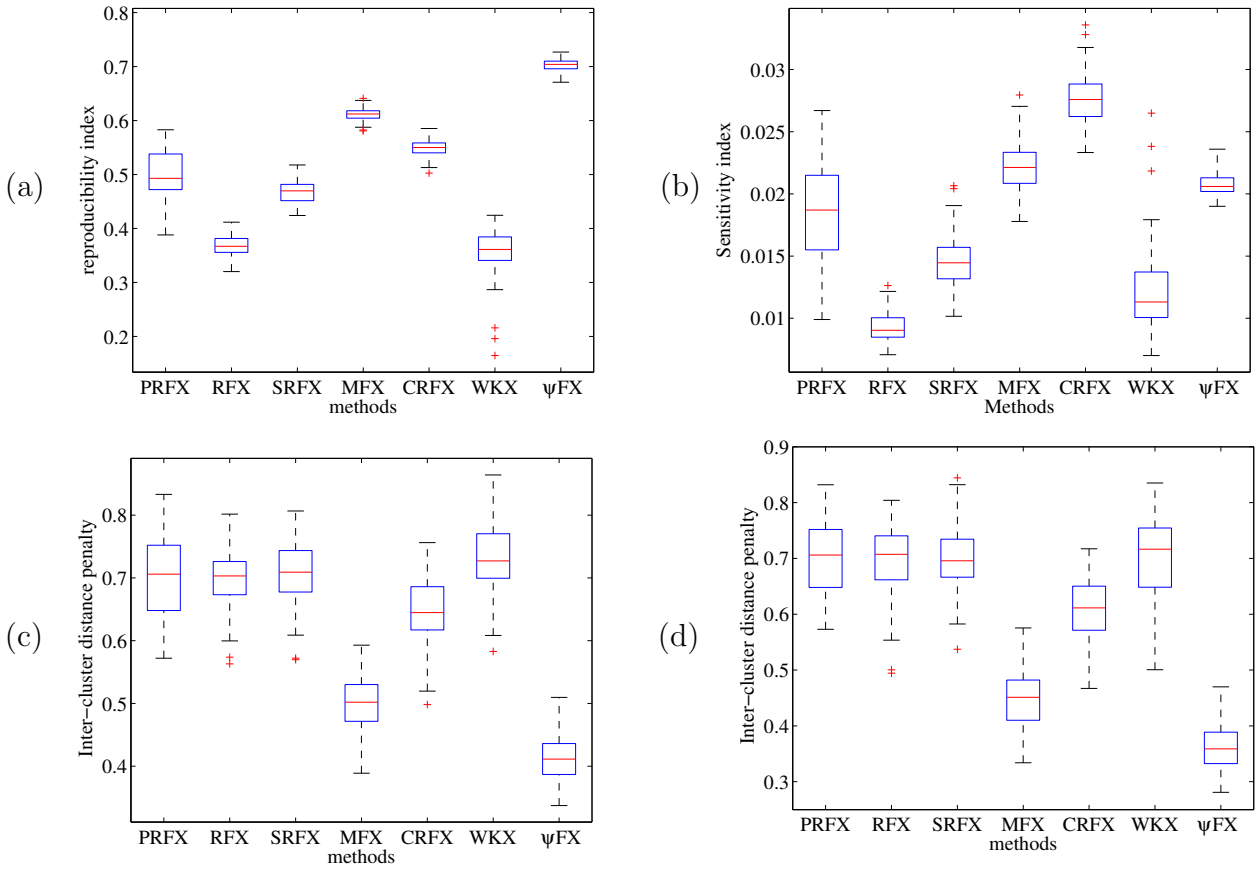


Figure 9: Dependence of the reliability κ (a), the sensitivity λ (b), inter-supra-threshold cluster distance penalty Φ (c-d) of the statistical analysis on the group statistic used. Φ is based clusters of size greater than $\eta = 10$ (c) or $\eta = 30$ (d). These quantities assessed considering $R = 8$ disjoint groups of size $\mathcal{S} = 10$ within the population of 81 subjects, using the *left click-right click* contrast.

3.5 Comparison of different group analysis methods

Now we study how the reliability index behaves for different statistical methods: The voxel-based RFX, the RFX test after 12 mm smoothing of the data -instead of 5mm- (SRFX), the MFX test, the parcel-based RFX test (PRFX), the cluster-level thresholded RFX (CRFX), The Wilkoxon test (WKX), and pseudo-MFX test ΨFX. RFX, SRFX, MFX, WKX, PRFX and ΨFX maps are thresholded at the $p < 0.001$ level, uncorrected for multiple comparisons. The CRFX map is first thresholded at the $p < 0.01$, uncorrected level, then at the $p < 0.01$ cluster-level. The results are obtained by bootstrapping in groups of size $\mathcal{S} = 10$. The results are presented in Fig. 9.

From the point of view of reliability, the WKX and RFX tests have the worst performance overall, while the SRFX performs slightly better. CRFX, PRFX and MFX techniques yield higher reliability, but ΨFX yield the highest values. The results are more variable with PRFX than with other techniques; this reflects the fact that PRFX is based on a smaller number of volume elements, so that statistical tests have a more crispy behaviour.

CRFX, MFX, and to a lesser extent, PRFX tests are more sensitive, i.e. have a larger fraction

of generally activated voxels, than voxel-based tests.

Finally, the average supra-threshold cluster distance Φ is minimal for ΨFX, and relatively low for MFX. It is approximately similar for the other techniques.

4 Discussion

4.1 Normality and Second-level variance

From Figures 3-5, one of the most striking effects is the co-localization of high second-level variance areas with high RFX statistic areas. Numerically, such an effect is not expected since the RFX is defined as the quotient of the mean effect by its standard deviation. Its interpretation could be that 1) the contrast-to-noise ratio (CNR) of the BOLD effect is highly variable across subjects, and by definition this effect does not appear in non-activated areas 2) spatial mis-registration[2] implies that at a given voxel, i.e. a given position in MNI space, some subjects have activity while other subjects have no activity there, thus widening the signal distribution. For simple contrasts such as those used (left or right click, sentence listening), different cognitive strategies should be ruled out. This inflated variance effect certainly deserves more investigation, given its prominent effect on statistics (sensitivity and reliability): for instance, the ΨFX statistic -that does not take into account the group variance, hence is simply a weighted average of the subject-based effects- seems more reliable than the MFX statistic, which is itself much more reliable than the RFX statistic (see Fig.9). The effect of group variance is also an argument in favour of Bayesian analysis of fMRI data, where the reference signal level is not 0 [Friston and Penny, 2003].

Non-normality is another important factor. To our knowledge, it has not been investigated earlier, since it requires a high number of subjects. Interestingly, the importance of non-normality is minored when considering normalized effects $\tau(s, v)$ instead of raw effects $\hat{\beta}(s, v)$. This confirms that first-level statistics can play an import role on group statistics. Interestingly, several areas with significant non-normality are found at the periphery of activation maxima, confirming the impact of spatial shifts in group statistics. once again, further investigations on non-normality may be performed, e.g. searching different groups of subject in the population or outlier subjects (see [Kherif et al., 2004]). The effect of non-normality can be evaluated by comparing the theoretical thresholds, based on normality assumption, with those derived non-parametrically, by random sign swap of the effects across subjects. In general, it is advisable to use non-parametric assessment to obtain reliable thresholds [Mériaux et al., 2006a]. However, the choice of the statistic should not necessarily be based on the avoidance of normal noise model: for instance, the Wilkoxon statistic, that adapts to non-normal data, did not perform better than other statistics in our experiments (see Fig. 9).

4.2 Measuring the reliability of group studies

The reliability of an activation pattern measures how systematically a given voxel or region will be found when performing a group study in one or other group of subjects. Taking advantage of

[2]Spatial mis-registration may be artefactual (incorrect normalization) or not (intrinsically different functional anatomy.)

the great number of subjects, we have used a bootstrap procedure, and two measures for assessing the reliability of the group studies: one models the activated/non-activated state of locations (voxel) as a mixture of binomial distributions, and quantifies the difference between the null and the active mode, while the other defines how well clusters of supra-threshold activity match across groups.

The first criterion has already been proposed in the literature ; it has the advantage of yielding very stable results across splits (see Figs. 6, 9); one reason is that all the R groups are used in each single computation of the parameters, while the cluster-based measure is based on pairwise comparisons. However, care must be taken because the estimation may be trapped in local minima (although we have never observed convergence problems in our experiments), or because the joint estimation of the different parameters may imply some non-trivial interaction between the parameters (e.g. the sensitivity λ might not be independent from κ). More importantly, results at the voxel level are not as important as the presence of a strong local maximum or a significant cluster, which deserve being reported.

This has incited use to develop a second measure (see Eq. (9)), which takes into account only extended clusters and compares the position of their centres. Note that the penalty function Φ stabilizes to $\Phi \simeq 1$ as soon as the distance exceeds 12mm.This is because clusters distant from e.g. 20 mm cannot be considered as homologous, as well as clusters distant from e.g. 50 mm (this is true because we are reporting group results; when reporting individual results, greater variability might be allowed). Averaging across supra-threshold clusters yields an idea of how frequently close clusters will be obtained across groups of subjects. This pairwise measure is somewhat more variable than the voxel-based indexes, but it yields an independent confirmation of possible differences in reliability.

As we have noticed, the dependence on the contrast studied is high with respect to the bootstrap dispersion, confirming the worthiness of these measures (Fig. 6). In general terms, κ and Φ have a similar behaviour (which corresponds to opposite numerical fluctuations: κ is high when Φ is low and vice versa). This was not obvious, given that the two measures are independent and based on completely different approaches. It suggests that our results are not artefactual, but really intrinsic to the data.

Our setting for the study of the reliability may also be used to compare competing pre-processing techniques or analysis frameworks, in addition to previous contributions based on cross-validation [Strother et al., 2002] and information theory [Kjems et al., 2002].

4.3 Is there any best threshold ?

The fundamental question of finding an optimal threshold to label areas as activated has rarely been tackled, since it requires the modelling at the voxel level of both the null and the converse hypothesis to control both the false positive and false negative rates. This is possible here, thanks to the high number of subjects. Interestingly, we find a relatively low value for the optimal threshold ($\theta^\star \sim 2.7$ when considering κ, $\theta^\star \sim 3$ when considering Φ ; note that these two measures are independent). The corresponding P-values ($0.0035 - 0.001$) are not conservative, so that such thresholds do not allow a very strict control of the rate of false positives. Family-wise error control procedures such as Bonferroni, Random Field Theory [Ashburner et al., 2004], and, to a lesser extent, False Discovery Rate [Genovese et al., 2002], typically imply the use of much higher

thresholds. In this study, we have chosen a relatively lenient threshold $p < 10^{-3}$ uncorrected because specificity control was not our main point. However, a good compromise between the control of false positives and the reliability may be the use of cluster-level or parcel-level inference.

We obtained very similar results with functional contrasts, like the auditory contrast, that are stronger in terms of signal, the only difference being that the optimal threshold was slightly higher, between 3 and 3.5. However, it is not obvious that our results generalize to datasets with different structure and our point is certainly not justify lenient thresholding procedures. Nevertheless, the question of an optimized threshold should be addressed more systematically in neuroimaging studies.

4.4 How does the sample size affect the reliability of the results ?

Another fundamental question concerns the number of subjects that should be included in a study. The point here is not only the sensitivity [Desmond and Glover, 2002, McNamee and Lazar, 2004], but also the reliability [Murphy and Garavan, 2004] of the results. Our results clearly indicate that $\mathcal{S} = 20$ is it a minimum if one wants to have acceptable reliability. As far as we know, most studies currently do not have this number of subjects, and one might be concerned with the reliability of many *findings* from neuroimaging studies.

One might object that in our case, only one session was available for each subject, and that the quick event-related design might yield poor results in terms of detection. However, the results that we describe are related to very basic contrasts (auditory and motor activity) for which we could check that most of the subjects (motor contrast) or even every subject (auditory contrast) had significant functional activity in expected regions, which has to be compared with the subtle functional contrasts that are currently topics of investigation. Moreover, our finding confirms earlier simulations and studies [Desmond and Glover, 2002, Murphy and Garavan, 2004]. For these reasons, we think that this paper should be considered as an incentive to include more subjects in neuroimaging studies.

4.5 Reliability of the different statistical tests

One of the most important practical questions is to describe or design the most efficient ways to perform group studies in neuroimaging. Our experiments are not exhaustive. In particular, some other techniques could have been introduced, as well as combination of techniques used in this work. But our first aim is to find general guidelines.

First of all, we think that non-parametric assessment of functional activity should be preferred to analytical tests, which rely on unvalidated and sometimes wrong hypotheses. This can be done using adapted toolboxes e.g. SnPM [Hayasaka and Nichols, 2003] or distance [Mériaux et al., 2006a]. It is worthwhile to note that the implementation of the tests in C reduces computation time to very reasonable time (cluster-level P-values can e.g. be computed in less than one minute on a dataset of ten subjects). Non-parametric estimation of the significance benefits both to the sensitivity and to the reproducibility of the studies.

Second, Mixed-Effects models should systematically be preferred to mere random effects analyses: there is some information in the first level of the data that improves the estimation of the group effects/variance and statistic.

Third, cluster- and parcel-based inference should be preferred to voxel-based thresholding. Cluster-level inference is of frequent use, which benefits to the sensitivity and the reliability of group analyses. However, it is based on the assumption that activated regions are large, which is not necessarily true. Parcel-based inference may thus be a interesting alternative, since it further allows some spatial relaxation in the subject-to-subject correspondence. The price to pay is a larger variability of the results due to a more crispy decision function (activated vs no activated). We recommend the combination of one of these techniques together with MFX. By contrast, stronger smoothing did not increase significantly the reliability of the results.

Fourth, to our surprise, ΨFX was found to be the most reliable technique. Although the statistic function does not take into account the group variance - as argued earlier, this is probably the reason of its higher performance - its distribution under the null hypothesis is tabulated by random swaps of the effects signs, so that it is indeed a valid group inference technique. However, care should be taken when using it: the thresholds have to be computed voxel per voxel, i.e. are not spatially stationary, second, the statistic value itself has no obvious interpretation, by contrast with the RFX or MFX statistics.

4.6 Conclusion

This analysis is a also start point to develop new strategies for brain mapping. Several directions may be addressed in the future:

- First trying to relate inter-subject variability to behavioral differences and individual or psychological characteristics of the subjects. Once again, such investigation may be undertaken only on large databases of subjects, and we the data basis used in this experiment might and probably will be used in such a framework.

- Second, efforts will further be made to relate spatial functional variability to anatomical variability. While some cortex-based analysis reports have indicated a greater sensitivity than standard volume-based mappings [Fischl et al., 1999], statistical evidence is still lacking, and it is not clear at all how much can be gained when taking into account macro-anatomical features, e.g. sulco-gyral anatomy. Similarly, diffusion-based imaging may add useful information to improve cross-subject brain cartography [Behrens et al., 2006].

- Third, at a statistical level, we think that intermediate levels of descriptions could be used more systematically between the subjects and the group level. Looking for outlier subjects, possible subgroups and so on can be investigated [Kherif et al., 2004, Thirion et al., 2005, Thirion et al., 2006], though finding a meaningful distance and separation criteria is not straightforward. For instance, it would be interesting to know what proportion of subjects had a significant activity in a given region; such a simple question requires to solve issues in across-subjects correspondences and in statistical thresholding (how can one be sure that two foci of activity in two subjects are homologous ?).

Finally, we hope that these procedures and results will provide useful guidelines in the experimental procedures and analysis of group functional neuroimaging datasets.

References

[Ashburner et al., 2004] Ashburner, J., Friston, K., and Penny, W., editors (2004). *Human Brain Function, 2nd Edition*. Academic press.

[Beckmann et al., 2003] Beckmann, C., Jenkinson, M., and Smith, S. (2003). General multi-level linear modelling for group analysis in fMRI. *Neuroimage*, 20:1052–1063.

[Behrens et al., 2006] Behrens, T. E. J., Jenkinson, M., Robson, M. D., Smith, S. M., and Johansen-Berg, H. (2006). A consistent relationship between local white matter architecture and functional specialisation in medial frontal cortex. *Neuroimage*, 30(1):220–227.

[Brammer et al., 1997] Brammer, M., Bullmore, E., Simmons, A., Grasby, P., Howard, R., Woodruff, P., and Rabe-Hesketh, S. (1997). Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach. *Magn. Reson. Imaging*, 15(7):763–770.

[Bullmore et al., 1999] Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. (1999). Global, voxel, and cluster tests, by theory and permutation, for difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.*, 18:32–42.

[Collins et al., 1998] Collins, D. L., G., L. G., and Evans, A. C. (1998). Non-linear cerebral registration with sulcal constraints. In *MICCAI'98, LNCS-1496*, pages 974–984.

[Desmond and Glover, 2002] Desmond, J. E. and Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J Neurosci Methods*, 118(2):115–128.

[Fischl et al., 1999] Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp*, 8(4):272–284.

[Flandin et al., 2002] Flandin, G., Kherif, F., Pennec, X., Malandain, G., Ayache, N., and Poline, J.-B. (2002). Improved detection sensitivity of functional MRI data using a brain parcellation technique. In *Proc. 5th MICCAI*, LNCS 2488 (Part I), pages 467–474, Tokyo, Japan. Springer Verlag.

[Friston et al., 2002] Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: Theory. *Neuroimage*, 16(2):465–483.

[Friston et al., 1999] Friston, K. J., Holmes, A. P., and Worsley, K. J. (1999). How many subjects constitute a study? *Neuroimage*, 10(1):1–5.

[Friston and Penny, 2003] Friston, K. J. and Penny, W. (2003). Posterior probability maps and SPMs. *Neuroimage*, 19(3):1240–1249.

[Genovese et al., 2002] Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

[Genovese et al., 1997] Genovese, C. R., Noll, D. C., and Eddy, W. F. (1997). Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. *Magn Reson Med*, 38(3):497–507.

[Hayasaka and Nichols, 2003] Hayasaka, S. and Nichols, T. (2003). Validating Cluster Size Inference: Random Field and Permutation Methods. *Neuroimage*, 20(4):2343–2356.

[Hellier et al., 2003] Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher., G., Collins, D. L., Evans, A., Malandain, G., Ayache, N., Christensen, G. E., and Johnson, H. J. (2003). Retrospective evaluation of intersubject brain registration. *IEEE Trans. Med. Imag.*, 22(9):1120–1130.

[Hollander and Wolfe, 1999] Hollander, M. and Wolfe, D. (1999). *Nonparametric statistical inference*. John Wiley & Sons, New York, USA, second edition edition.

[Holmes et al., 1996] Holmes, A., Blair, R., Watson, J., and Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.*, 16:7–22.

[Jernigan et al., 2003] Jernigan, T. L., Gamst, A. C., Fennema-Notestine, C., and Ostergaard, A. L. (2003). More "mapping" in brain mapping: statistical comparison of effects. *Hum Brain Mapp*, 19(2):90–5.

[Kherif et al., 2004] Kherif, F., Poline, J.-B., Mériaux, S., Benali, H., Flandin, G., and Brett, M. (2004). Group analysis in functional neuroimaging: selecting subjects using similarity measures. *Neuroimage*, 20(4):2197–2208.

[Kjems et al., 2002] Kjems, U., Hansen, L. K., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., and Strother, S. C. (2002). The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. *Neuroimage*, 15(4):772–786.

[LaConte et al., 2003] LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., and Strother, S. (2003). The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. *Neuroimage*, 18(1):10–27.

[Liou et al., 2005] Liou, M., Su, H.-R., Lee, J.-D., Aston, J. A. D., Tsai, A. C., and Cheng, P. E. (2005). A method for generating reproducible evidence in fMRI studies. *Neuroimage*.

[Liou et al., 2003] Liou, M., Su, H.-R., Lee, J.-D., Cheng, P. E., C.-C., H., and Tsai, C.-H. (2003). Bridging functional MR images and scientific inference: Reproducibility maps. *Journal of Cognitive Neuroscience*, 15(7):935–945.

[Maitra et al., 2002] Maitra, R., Roys, S. R., and Gullapalli, R. P. (2002). Test-retest reliability estimation of functional MRI data. *mrm*, 48(1):62–70.

[McNamee and Lazar, 2004] McNamee, R. L. and Lazar, N. A. (2004). Assessing the sensitivity of fMRI group maps. *Neuroimage*, 22(2):920–931.

[Mériaux et al., 2006a] Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., and Poline, J.-B. (2006a). Combined permutation test and mixed-effect model for group average analysis in fMRI. *Hum. Brain Mapp.*, pages 402–410.

[Mériaux et al., 2006b] Mériaux, S., Roche, A., Thirion, B., and Dehaene-Lambertz, G. (2006b). Robust statistics for nonparametric group analysis in fMRI. In *Proc. 3th Proc. IEEE ISBI*, pages –, Arlington, VA.

[Murphy and Garavan, 2004] Murphy, K. and Garavan, H. (2004). An empirical investigation into the number of subjects required for an event-related fMRI study. *Neuroimage*, 22(2):879–85.

[Neumann and Lohmann, 2003] Neumann, J. and Lohmann, G. (2003). Bayesian second-level analysis of functional magnetic resonance images. *Neuroimage*, 20(2):1346–1355.

[Nichols and Holmes, 2002] Nichols, T. and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.*, 15:1–25.

[Shaw et al., 2003] Shaw, M. E., Strother, S. C., Gavrilescu, M., Podzebenko, K., Waites, A., Watson, J., Anderson, J., Jackson, G., and Egan, G. (2003). Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. *Neuroimage*, 19(3):988–1001.

[Smith et al., 2005] Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. M., and McGonigle, D. J. (2005). Variability in fMRI: a re-examination of inter-session differences. *Hum Brain Mapp*, 24(3):248–57.

[Stiers et al., 2006] Stiers, P., Peeters, R., Lagae, L., Hecke, P. V., and Sunaert, S. (2006). Mapping multiple visual areas in the human brain with a short fMRI sequence. *Neuroimage*, 29(1):74–89.

[Strother et al., 2002] Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage*, 15(4):747–71.

[Thirion et al., ress] Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., and Poline, J.-B. (2005, in Press). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.*

[Thirion et al., 2005] Thirion, B., Pinel, P., and Poline, J.-B. (2005). Finding landmarks in the functional brain: Detection and use for group characterization. In *Proc. MICCAI2005*, Palm Spings, USA.

[Thirion et al., 2006] Thirion, B., Roche, A., Ciuciu, P., and Poline, J.-B. (2006). Improving sensitivity and reliability of fmri group studies through high level combination of individual subjects results. In *Proc. MMBIA2006*, New York, USA.

[Wei et al., 2004] Wei, X., Yoo, S.-S., Dickey, C. C., Zou, K. H., Guttmann, C. R. G., and Panych, L. P. (2004). Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *Neuroimage*, 21(3):1000–8.

[Woolrich et al., 2004] Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. (2004). Multi-level linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747.

[Worsley et al., 2002] Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fMRI data. *Neuroimage*, 15(1):1–15.

[Worsley, 2005] Worsley, K. J. (2005). An improved theoretical P value for SPMs based on discrete local maxima. *Neuroimage*, 28(4):1056–1062.

[Zar, 1999] Zar, J. H. (1999). *Biostatistical Analysis*. Prentice-Hall, Inc., Upper Saddle River, NJ.