NeuroImage 284 (2023) 120457

Contents lists available at ScienceDirect

NeuroImage

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

The impact of emotional valence and stimulus habituation on fMRI signal reliability during emotion generation

Alice Pirastru[a,b], Sonia Di Tella[a,c], Marta Cazzoli[a], Fabrizio Esposito[d], Giuseppe Baselli[b], Francesca Baglio[a,*], Valeria Blasi[a]

[a] IRCCS Fondazione Don Carlo Gnocchi, ONLUS, Milan, Italy
[b] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
[c] Department of Psychology, Università Cattolica del Sacro Cuore, Milan, Italy
[d] Department of Advanced Medical and Surgical Sciences, University of Campania "Luigi Vanvitelli", Naples, Italy

ARTICLE INFO

Keywords:
fMRI
Rehabilitation
Repetition suppression
Signal reliability
Test-retest reliability
Emotion generation

ABSTRACT

Background: The emotional domain is often impaired across many neurological diseases, for this reason it represents a relevant target of rehabilitation interventions. Functional changes in neural activity related to treatment can be assessed with functional MRI (fMRI) using emotion-generation tasks in longitudinal settings. Previous studies demonstrated that within-subject fMRI signal reliability can be affected by several factors such as repetition suppression, type of task and brain anatomy. However, the differential role of repetition suppression and emotional valence of the stimuli on the fMRI signal reliability and reproducibility during an emotion-generation task involving the vision of emotional pictures is yet to be determined.
Methods: Sixty-two healthy subjects were enrolled and split into two groups: group A (21 subjects, test-retest reliability on same-day and with same-task-form), group B (30 subjects, test-retest reproducibility with 4-month-interval using two equivalent-parallel forms of the task). Test-retest reliability and reproducibility of fMRI responses and patterns were evaluated separately for positive and negative emotional valence conditions in both groups. The analyses were performed voxel-wise, using the general linear model (GLM), and via a region-of-interest (ROI)-based approach, by computing the intra-class correlation coefficient (ICC) on the obtained contrasts.
Results: The voxel-wise GLM test yielded no significant differences for both conditions in reliability and reproducibility analyses. As to the ROI-based approach, across all areas with significant main effects of the stimuli, the reliability, as measured with ICC, was poor (<0.4) for the positive condition and ranged from poor to excellent (0.4–0.75) for the negative condition. The ICC-based reproducibility analysis, related to the comparison of two different parallel forms, yielded similar results.
Discussion: The voxel-wise GLM analysis failed to capture the poor reliability of fMRI signal which was instead highlighted using the ROI-based ICC analysis. The latter showed higher signal reliability for negative valence stimuli with respect to positive ones. The implementation of two parallel forms allowed to exclude neural suppression as the predominant effect causing low signal reliability, which could be instead ascribed to the employment of different neural strategies to cope with emotional stimuli over time. This is an invaluable information for a better assessment of treatment and rehabilitation effects in longitudinal studies of emotional neural processing.

1. Introduction

The emotional domain has proven to be affected by several neurological disorders such as dementia (Bora et al., 2016; Bora and Yener, 2017; Klein-Koerkamp et al., 2012) and Parkinson's Disease (Anzuino et al., 2023; Blonder and Slevin, 2011; Gray and Tickle-Degnen, 2010). This domain is often targeted during rehabilitation interventions, and the effects of the treatment at the neural level can be assessed through the implementation of specific emotion-generation tasks using functional magnetic resonance imaging (fMRI).

[*] Corresponding author.
E-mail address: fbaglio@dongnocchi.it (F. Baglio).

https://doi.org/10.1016/j.neuroimage.2023.120457
Received 28 July 2023; Received in revised form 9 November 2023; Accepted 13 November 2023
Available online 15 November 2023

Indeed, fMRI represents the gold standard among non-invasive neuroimaging techniques to assess functional reorganization consequent to a neurorehabilitation intervention. Given its high spatial sensitivity, task-based fMRI has been widely used for deriving robust functional activation markers, which can effectively investigate disease evolution and treatment effects in longitudinal studies (Drobyshevsky et al., 2006). However, the test-retest assessment of signal reliability of functional MRI measures has been a matter of debate for the last decades (Elliott et al., 2020; "Fostering reproducible fMRI research," 2017; Noble et al., 2021). Indeed, the reliability of a measure is extremely important to uniquely describe a variable and a prerequisite for identifying novel imaging-derived biomarkers (Elliott et al., 2020). The reliability and reproducibility of the signal over repeated sessions is specifically fundamental when dealing with longitudinal measures assessing the progression of a given pathology or the effect of a treatment. Yet, as stated by Bennet and Miller (Bennett and Miller, 2010), a consensus about the fMRI signal reliability is still lacking.

Several factors can impact signal reliability, one of the main being the type of fMRI task. In fact, relevant sources of variability in the fMRI signal are often ascribed to the underlying mental strategies employed when performing a task. Different studies have highlighted heterogeneous results in terms of reliability depending on the type of task (Holiga et al., 2018), with sensory and motor tasks being more reliable than tasks involving higher cognitive processes (Bennett and Miller, 2010). Cognitive tasks can indeed rely upon different mental strategies, requiring the activation of different neuronal substrates, to respond to the same task over time (Bennett and Miller, 2010). The impact of the employed neural strategy is extremely relevant when considering an emotion stimulation task in which for instance subjects can decrease the perceived intensity of stimuli with negative valence by using reappraisal as a regulatory strategy (Berboth et al., 2021). Interestingly, this variability has been ascribed to increased neural efficiency in emotion processing associated with a decreased cognitive effort over time. Moreover, in test-retest assessment, the reiteration of task stimuli can be associated with the phenomenon of repetition suppression (Grill-Spector et al., 2006; Segaert et al., 2013) also known as fMRI adaptation, that is, the reduction of neural activity and Blood oxygen level-dependent (BOLD) signal when stimuli are repeated. This phenomenon can occur at different temporal scales, ranging from minutes to months, and across multiple brain regions. Other sources of variability, affecting signal reliability have been identified, such as the higher SNR characterizing cortical brain regions with respect to subcortical ones (Heilicher et al., 2022). An example of these is given in (McDermott et al., 2020), showing how the reliability of a visual-cued emotion-processing task decreased from visual sensory areas to limbic ones. Lately, Berboth et al. (Berboth et al., 2021) assessed the fMRI signal reliability in an emotion regulation task involving visual negative stimuli, within four networks of interest derived from a recent meta-analysis (Morawetz et al., 2020), highlighting great variability in reliability measures depending on the observed brain areas.

Altogether these findings seem to indicate that different factors may affect the reliability of the fMRI signal, especially considering complex, higher-order mental processes such as emotion processing. Specifically related to emotion stimulation tasks, the observed variability has been ascribed to increased neural efficiency due to the use of different neural strategies and brain circuits over time, while the role of the repetition suppression phenomenon due to a habituation effect to salient stimuli has not been specifically addressed.

Finally, even though emotional valence (i.e., positive or negative) could elicit different coping strategies and therefore engage different neural circuits over time, its role in signal reliability has never been investigated.

To summarise, both the repetition suppression and the use of different mental strategies are highly relevant in the context of longitudinal studies where the possibility to disentangle between these and the actual effects related to the evolution of a clinical condition or to a specific treatment is pivotal. However, studies investigating both these aspects simultaneously and assessing their effect on fMRI signal reliability are lacking.

In this framework, the present study aimed to assess the impact of repetition suppression and of different emotional valence on the fMRI signal reliability during an emotion stimulation task. To differentiate the role of each of these mechanisms in fMRI, we evaluated test-retest signal reliability of an emotion generation task based on the vision of pictures with emotional valence, and signal reproducibility between two parallel forms of the same task specifically selected to reduce stimulus repetition in two different experimental conditions involving stimuli with positive and negative valence. Considering that signal reliability and reproducibility were also tested according to different time scales (e.g. within the same day and with an average interscan interval of 4-months respectively), we expected the employment of the parallel forms to be effective for improving within subject signal reliability if the source of variability can be mainly ascribed to repetition suppression rather than to the changes in mental strategies and modulation implemented at the individual level. On the other hand, if the variability is instead mainly linked to the over-time changes of the employed neural strategy, we predicted to observe different degrees of reliability also depending on the experimental condition (i.e., the different emotional valence of the stimuli).

2. Methods

2.1. Participants

Sixty-two right-handed healthy volunteers were enrolled in the study (mean age ± standard deviation in years= 27.2 ± 7.9; 35 females; mean education ± standard deviation in years 16.2 ± 2.2). The absence of neurological, and neuropsychiatric disorders, and the use of psychotropic drugs were assessed through a clinical interview and considered as inclusion criteria. The exclusion criteria consisted of any contraindication to perform an MRI examination (e.g., presence of metallic prosthetics, MRI unsafe devices, claustrophobia, etc.). All the subjects signed a written informed consent, and the study was approved by the IRCCS Fondazione Don Gnocchi Ethical Committee.

At the recruitment (T0), subjects were randomly allocated to one of the two parallel forms of the same task (form 1 and form 2). After the first fMRI scan, subjects were split into two different groups, group A and group B composed respectively of 1/3 and 2/3 of the total sample size (see Fig. 1). A higher sample size for group B was considered to account for possible drop-outs due to the extended inter-scan interval. The subjects belonging to group A repeated the acquisition twice within the same day, and with the same parallel form (form 1) to assess the test-retest reliability. The subjects belonging to group B instead repeated the acquisition twice on different days and considering different parallel forms of the task (i.e., either T0= form 1 – T1= form 2 or T0= form 2 – T1= form 1), with an average interscan interval of 4 months (T1), to test for the reproducibility (equivalence) of the two parallel forms (form 1 and form 2). The employment of different stimuli in the two parallel forms and the prolonged inter-scan interval ensured the removal of potential repetition suppression phenomena.

Eleven subjects, belonging to group B, did not come back for the second MRI scanning session; the resulting sample used for test-retest reliability and reproducibility analysis was composed of 51 subjects: 21 subjects belonging to group A (mean age ± standard deviation in years= 25.9 ± 6; 8 females; mean education ± standard deviation in years 16.3 ± 1.9) and 30 subjects belonging to group B (mean age ± standard deviation in years= 27.1 ± 7.2; 21 females; mean education ± standard deviation in years 16.2 ± 2). The sample size and demographics are reported in Table 1.

2.2. Task design

The emotional stimulation paradigm consisted of a block design

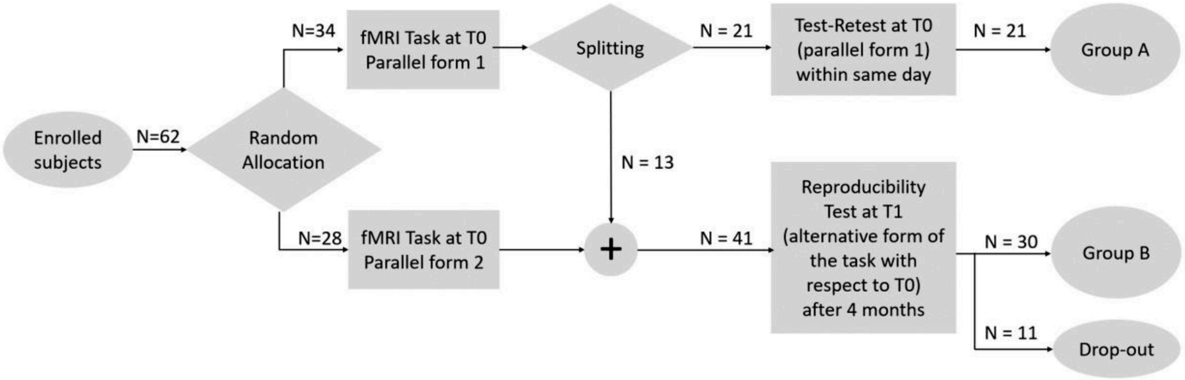

Enrolled subjects
N=62
Random Allocation
N=34
fMRI Task at T0 Parallel form 1
Splitting
N = 21
Test-Retest at T0 (parallel form 1) within same day
N = 21
Group A
N = 13
N=28
fMRI Task at T0 Parallel form 2
+
N = 41
Reproducibility Test at T1 (alternative form of the task with respect to T0) after 4 months
N = 30
Group B
N = 11
Drop-out

Fig. 1. Flow-chart of the subjects' enrolment and allocation. T0= enrolment; T1= average interscan interval of 4 months.

Table 1
Sample size, groups and demographics. Legend: *N*= sample size; SD= standard deviation; ROI= region of interest.

Group (N)	Whole sample (62)	Group A (21)	Group B (30)
Age in years, mean ± SD	27.2 ± 7.9	25.9 ± 6	27.1 ± 7.2
Females, number (%)	35 (56 %)	8 (38 %)	21 (70 %)
Education in years, mean ± SD	16.2 ± 2.2	16.3 ± 1.9	16.2 ± 2
Parallel Form	1 and 2	1	1 and 2
Type of Analysis	Task-related pattern of activation (voxel-wise)	Test-Retest signal reliability (voxel-wise and ROI-based)	Signal reproducibility in parallel forms (voxel-wise and ROI-based)

comprising 4 blocks representing 4 different conditions (A-B-C-D, see Fig. 2) which were repeated 8 times (epochs): presentation of visual stimuli with positive (A), neutral (B), and negative (C) emotional valence and fixation of a white cross (D). In order to ensure that participants were visualizing the images, a simple, not cognitively demanding task, was designed. Specifically, subjects were asked to indicate, by pressing a button with either right index or middle finger, if the presented scene included either living or non-living elements respectively, as previously implemented in (Pfaff et al., 2019). All the subjects received a proper training using trial images before the scanning.

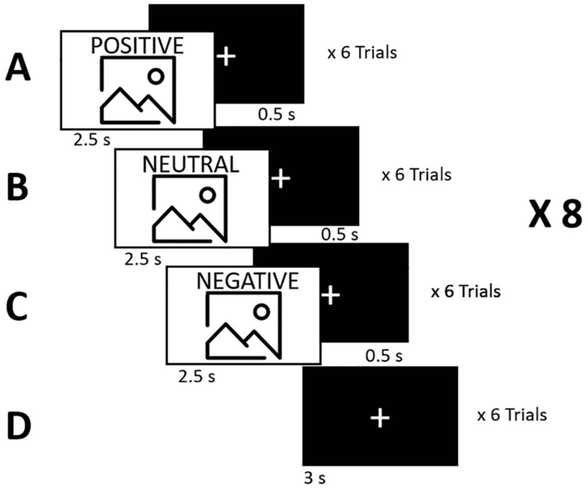


Fig. 2. Emotion stimulation task paradigm. The paradigm comprises 4 blocks representing 4 different conditions (A-B-C-D): presentation of visual stimuli with positive (A), neutral (B) and negative (C) emotional valence and fixation of a white cross (D), which are repeated 8 times (8 epochs). Every block consists of 6 different trials all of the same emotional valence which were displayed for 2.5 s each and interleaved by a white cross fixation (0.5 s), for a total duration of 18 s. The fixation condition was implemented to last 18 s. The presentation of the blocks was randomized for each epoch.

Each stimulus was presented only once within each scanning session. The visual stimuli were selected from the International Affective Picture System (IAPS) database (Lang et al., 2008). Specifically, the stimuli were selected according to the average scoring reported for emotional valence (mean ± SD form 1: 7.45 ± 0.37 for positive, 2.68 ± 0.77 for negative, and 5.08 ± 0.57 for neutral; form 2: 7.45 ± 0.43 for positive, 2.82 ± 0.79 for negative, 5.20 ± 0.61 for neutral), arousal (mean ± SD form 1: 5.27 ± 0.68 for positive, 5.31 ± 0.89 for negative, and 2.82 ± 0.39 for neutral, form 2: 5.30 ± 0.7 for positive, 5.42 ± 0.98 for negative, and 2.86 ± 0.42 for neutral) and dominance (mean ± SD form 1: 6.12 ± 0.57 for positive, 3.94 ± 0.8 for negative, and 6.04 ± 0.49 for neutral, form 2: 6.08 ± 0.5 for positive, 3.81 ± 0.78 for negative, and 5.99 ± 0.42 for neutral).

The blocks were presented in a randomised order within each epoch. Every block consisted of 6 different images (trials) all of the same emotional valence which were displayed for 2.5 s each and interleaved by a white cross fixation (0.5 s), for a total duration of 18 s. The fixation condition was implemented in order to last 18 s as well. Every epoch, comprising 4 blocks, lasted 72 s. The total scanning time of the task was approximately 11 min.

2.3. Parallel forms implementation

The parallel forms were created by matching the selected stimuli according to their valence, arousal and dominance derived from the normative data rating (Lang et al., 2008). Ninety-six total images per emotional valence were selected to construct 2 parallel forms of the experiment comprising 48 images per emotional valence each. Specifically, the two sets of images were attentively controlled to be as matched as possible concerning the indexes of valence, arousal, dominance, for positive, negative and neutral stimuli respectively and across emotional valence within the same task form. Independent sample t-test and Mann-Whitney test yielded no significant differences between the two parallel forms relatively to any of the considered indexes. Furthermore, we considered the semantic category, namely the prevalence of human subjects, animals or inanimate objects in the images, as an additional feature for stimulus matching both between parallel forms and across stimuli with different emotional valence.

The detailed method and results of the parallel forms implementation are reported in the Supplementary Materials (Table S1, Table S2, Table S3).

2.4. MRI data acquisition and analysis

2.4.1. MRI data acquisition

MRI acquisition was performed on a 3T Siemens Prisma Scanner (Erlangen, Germany) equipped with a 64-channel head/neck coil. The protocol included: 1) a T1–3D magnetization prepared rapid acquisition with gradient-echo (MPRAGE) sequence with repetition time (TR)= 2300 ms, echo time (TE)= 3.1 ms, isotropic resolution= 0.8 × 0.8 × 0.8 mm^3, 224 slices, which was used as an anatomical reference; 2) a sagittal fluid attenuated inversion recovery (FLAIR) sequence was also acquired (TR = 5000 ms, TE = 394 ms, resolution = 0.8 × 0.8 × 1 mm^3, acquisition matrix = 288 × 320, 176 slices), to exclude gross brain abnormalities; 3) an accelerated GE sequence with TR= 2000 ms, TE= 30 ms, resolution 3 × 3 × 3 mm^3, multi-slice acceleration factor= 2, 52 slices, 330 measurements, which was acquired during the task administration.

The IAPS visual stimuli were delivered using E-Prime 3.0 (psychology software tools, https://pstnet.com/products/e-prime/) by means of a NordicNeuroLab system (https://www.nordicneurolab.com/) comprising an “in-room viewing device” with an MR-compatible display located at the end of the gantry and a mirror placed on the head coil. The stimuli administration was synchronised with the MR acquisition by means of a dedicated device (SyncBox).

2.4.2. MRI data preprocessing

The anatomical MPRAGE volumes were pre-processed following the steps of bias field correction (Tustison et al., 2010) and brain extraction (Jenkinson et al., 2005; Smith, 2002) and were used as anatomical reference in the co-registration steps of the functional image processing.

The fMRI analyses were performed using the Statistical Parametric Mapping toolbox (SPM12, https://www.fil.ion.ucl.ac.uk/spm/) according to a standard pipeline comprising the following steps: motion correction and realignment, co-registration with individuals’ anatomical volumes, segmentation and normalization to the standard MNI template and smoothing (8 mm full-width at half-maximum isotropic Gaussian). The degree of head motion was assessed and subjects with movements above the threshold set at 2 mm/2° were excluded from the analysis.

2.4.3. MRI statistics

The general linear model (GLM) was used to construct and fit the statistical model on the BOLD response to perform the first-level analysis. The four experimental conditions (positive, negative, neutral stimuli, and fixation) were considered as regressors of interest; the six motion parameters were instead inserted in the model as nuisance regressors. Two-different contrasts were derived at the subject level comparing the positive or negative blocks to the neutral blocks, namely testing the positive and negative contrasts respectively.

The activation maps were derived from the whole sample of subjects ($n = 62$) for both the positive and negative contrasts by means of GLM and one-sample *t*-tests. The functional maps were considered statistically significant for $p_{FWE} < 0.05$ considering the family-wise error (FWE) correction for multiple comparisons to account for false positives. A threshold on cluster size was also set to a minimum size of 30 voxels.

2.5. ROI definition

To perform test-retest reliability and reproducibility analyses the regions of interest (ROIs) were obtained from the intersection between activation clusters (as resulting from the voxel-based GLM analysis) and an a priori selection of brain parcels from a standard anatomical atlas. To minimise the selection bias in the ROI definition, a mixed approach was used considering both the actual activation of the whole sample of subjects included in the study and literature evidence. Specifically, the clusters of significant activations from our sample were mapped according to the automated anatomical labelling 3 (AAL3) atlas (Rolls et al., 2020), using the AAL3 toolbox embedded in SPM12, and the percentage of the overlap of the activation cluster with respect to the parcel size was calculated. The median value of the overlap between the activation cluster and the AAL3 parcel was used as a threshold to exclude clusters with negligible overlap relative to the parcel size. Then, the parcels were further refined according to a recent meta-analysis on emotion regulation tasks (Laird et al., 2009; Morawetz et al., 2020), from which the areas belonging to the networks involved in emotional responses and perception, and processing of internal sensations were selected.

2.6. Statistical analyses: test-retest repeatability and reproducibility of parallel forms

To assess firstly the test-retest reliability of the same version of the task and secondly the reproducibility (equivalence) of the two implemented parallel forms, either a voxel-wise GLM-based approach and an ROI-based analysis were employed.

A whole-brain voxel-wise analysis unavoidably suffers from low statistical power, since statistical tests are carried out across all single voxels. Nonetheless, it is the standard method employed to perform statistical group comparisons in the anatomically normalised space and to detect clusters of significant fMRI signal changes over time due to either treatment or rehabilitation effects.

For this reason, the test-retest reliability analysis was also implemented as a ROI-based analysis by computing the intra-class correlation coefficient (ICC). The ICC is a robust index assessing the test-retest reliability of any metric which relies on the ratio between inter-subject and intra-subject variances (Heilicher et al., 2022).

To assess the test-retest reliability the analyses were performed on the subsample of 21 subjects, namely group A; the reproducibility of the parallel forms was instead tested on the subsample comprising 30 subjects, namely group B.

The voxel-wise analyses were conducted using a GLM approach to extract the main stimulus contrasts at single subject level (separately for positive and negative emotion conditions) to compute paired t–tests, and, in a second-level (group) analysis, to assess both test-retest reliability and parallel forms reproducibility. The results were assessed considering pFWE < 0.05 with cluster size threshold equal to 30 voxels.

The ICC was used to assess the ROI-based test-retest reliability and the equivalence of the parallel forms. Namely, the individual contrasts for each condition and form were extracted from each ROI and a two-way mixed effects model with an absolute agreement, namely ICC (2,1) (McGraw and Wong, 1996), was computed using SPSS (IBM Corporation, version 28).

The reliability, as measured by ICC, has been qualitatively categorised according to the following ranges: poor ICC < 0.4, fair $0.4 \leq$ ICC < 0.59, good $0.6 \leq$ ICC ≤ 0.75, and excellent reliability ICC > 0.75 (Cicchetti and Sparrow, 1981). Despite ICC values being categorised for positive ranges, negative ICC values could be also obtained, especially in neuroimaging studies, in which case, these are either categorised as poor to zero reliability or considered as uninterpretable (Bartko, 1976; Giraudeau, 1996; Lahey et al., 1983).

3. Results

3.1. Task-related pattern of activation

The whole brain activity related to the different contrasts of the task was investigated considering the first acquisition (T0) of the whole sample ($n = 62$). The significant main effects for positive and negative contrasts are shown in Fig. 3, whereas cluster size, peak statistics and localization according to AAL atlas are reported in Table 2 and Table 3 respectively for positive and negative contrasts.

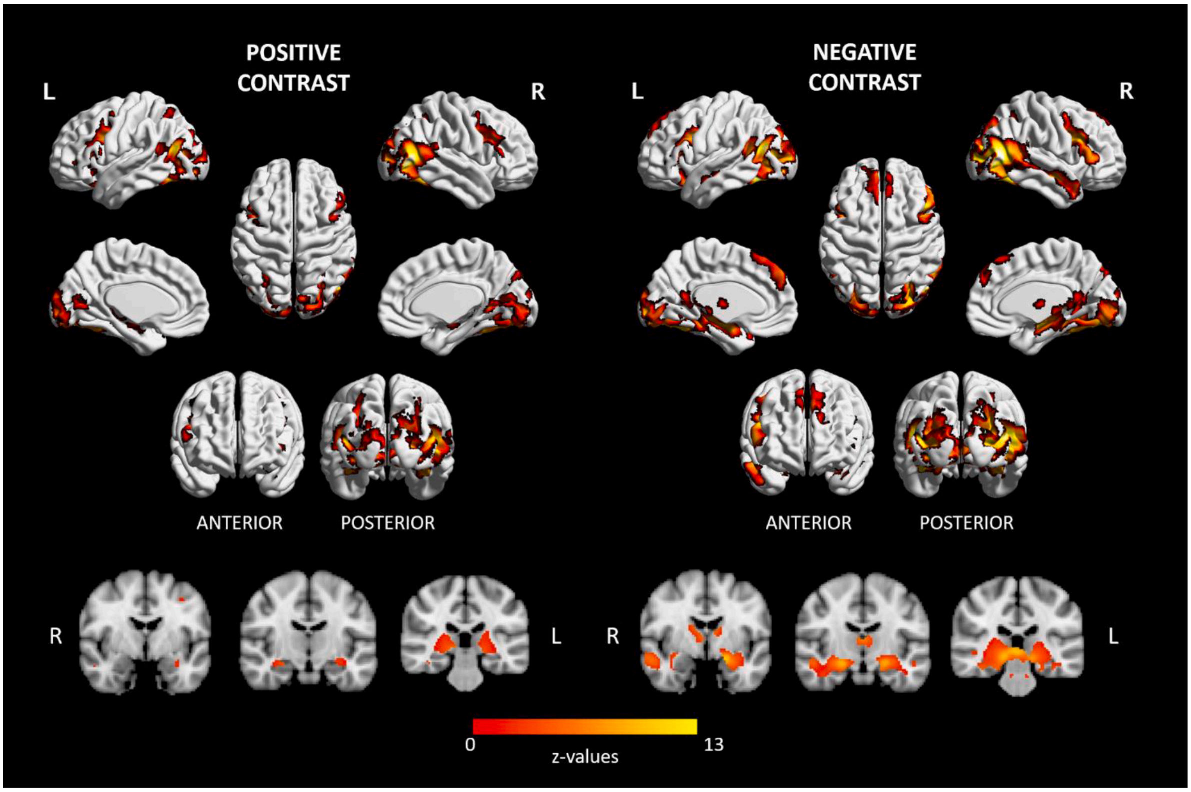


Fig. 3. Task elicited whole brain activity (pFWE < 0.05) for the positive (left) and negative (right) contrast. The cluster represents z-values color-coded in red-yellow according to the range reported in the colorbar (0–13). The glass-brain representation was created using the BrainNet Viewer Software (http://www.nitrc.org/projects/bnv/) (Xia et al., 2013). Legend: *R*= Right; *L*= Left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. *ROI selection*

For the positive condition, a total of 18 ROIs were selected while 22 ROIs were identified for the negative condition. The included ROIs are listed in Table 4.

3.3. *Test-Retest reliability analysis*

3.3.1. *Reliability - GLM voxel-wise analysis*

The within subject (paired *t*-test) between-session (Test vs. Retest) comparison for both Positive Condition and Negative Condition contrasts yielded no significant effects (pFWE < 0.05 at voxel-level).

3.3.2. *Reliability - Intra-class correlation coefficient – ROI-based analysis*

The test-retest ICC results for the positive and negative conditions are shown in Fig. 4.

For the Positive condition ICC ranged between 0.016 and 0.31 (poor reliability), while the Negative condition showed in general higher ICC values, spanning between 0.192 and 0.819 (poor to excellent reliability).

3.4. *Equivalence of parallel forms: reproducibility analysis*

3.4.1. *Reproducibility - GLM voxel-wise analysis*

No significant effects were detected when comparing the different parallel forms using a paired *t*-test for both the Positive and Negative conditions (pFWE < 0.05).

3.4.2. *Reproducibility - Intra-class correlation coefficient – ROI-based analysis*

The test-retest ICC results for the positive and negative conditions are shown in Fig. 5.

For the Positive condition ICC ranged between 0.005 and 0.479 (poor to fair reproducibility), with fewer negative values with respect to the reliability analysis. The Negative condition showed again higher ICC values than the Positive one but was more heterogeneous with respect to the reliability analysis, spanning between 0.03 and 0.695 (poor to good reproducibility).

4. Discussion

The present work aimed to investigate the impact of repetition suppression (i.e., habituation effect to repeated stimuli) and task conditions (i.e., different emotional valence) on fMRI signal reliability of an emotion generation task. This type of task investigates a domain which is often targeted during rehabilitation interventions for its involvement in several neurological disorders such as dementia (Bora et al., 2016; Bora and Yener, 2017; Klein-Koerkamp et al., 2012) and Parkinson's Disease (Anzuino et al., 2023; Blonder and Slevin, 2011; Gray and Tickle-Degnen, 2010). However, it is a common experience that facing emotions elicits mental strategies able to alter the nature, intensity, and duration of the neural (and behavioural) response.

In our study, the activation patterns observed at T0 in positive versus neutral and negative versus neutral contrasts encompassed a large bilateral network involving inferior-frontal areas, occipital-temporal areas, insular cortex, thalamus, amygdala, caudate nucleus and hippocampus in line with (Berboth et al., 2021).

This study addressed both the test-retest reliability of the task itself and the reproducibility of the two parallel forms implemented. Two different statistical approaches were used to assess the test-retest signal reliability and signal reproducibility, namely a standard GLM voxel-wise approach and an ROI-based approach which comprised the computation of the ICC (McGraw and Wong, 1996). The first approach has been included in our study since it represents a standard statistical method

Table 2
Task elicited whole brain activity for the positive condition. The activation clusters are reported considering the cluster size, the statistics in terms of both t-value and p-value, the peak of activation and their localization according to the AAL atlas. Legend: FWE= family wise error correction for multiple comparisons; equivk= cluster size. *R*= Right; *L*= Left; C= cortex.

p(FWE-corr)	equivk	T	x,y,z [mm]	AAL Label
<0.001	14,308	13.10	50 −70 6	R Middle Temporal C
		11.04	−46 −70 18	L Middle Temporal C
		11.15	22 −96 −4	R Calcarine C
		10.25	−38 −48 −20	L Fusiform
		11.54	42 −48 −20	R Fusiform
		11.39	26 −96 −2	R Inferior Occipital C
		8.25	−26 −94 −4	L Inferior Occipital C
		10.15	−44 −76 8	L Middle Occipital C
		9.73	44 −76 8	R Middle Occipital C
		10.01	20 −98 4	R Superior Occipital C
<0.001	557	9.53	−24 −30 2	L Thalamus
		5.65	−30 −4 −20	L Amygdala
		5.89	−20 −22 20	L Caudate
		7.37	−34 −12 −14	L Hippocampus
<0.001	344	9.38	24 −28 2	R Thalamus
<0.001	784	8.63	−40 10 30	L Precentral C
		6.10	−44 16 16	L Inferior Frontal Opercular C
<0.001	1111	7.37	52 20 32	R Inferior Frontal Opercular C
		6.74	42 4 40	R Middle Frontal C
		6.13	42 4 46	R Precentral C
<0.001	247	7.36	−22 −34 −44	Cerebellum_10_L
		5.41	−10 −54 −46	Cerebellum_9_L
		6.49	2 −56 −36	Vermis_9
<0.001	149	7.24	−46 38 0	L Inferior Frontal Triangular C
<0.001	111	6.91	−28 18 −22	L Inferior Orbitofrontal C
		5.89	−28 22 −2	L Insula
<0.001	81	6.16	32 −14 −14	R Hippocampus
0.002	33	6.56	22 −36 −46	Cerebrum_10_R
<0.001	123	6.14	34 −64 −48	Cerebellum_8_R

employed in both cross-sectional studies comparing different populations and longitudinal ones testing for treatment-related neural effects. However, we also included the ICC computation because it represents a more suitable methodology for quantifying the test-retest signal reliability (Bennett and Miller, 2010; Heilicher et al., 2022; Noble et al., 2021), relying on the ratio of between-subject and between-session variance (Shrout and Fleiss, 1979). The latter approach overcomes the limitation of the GLM approach which is intrinsically affected by low statistical power, because statistical tests are carried out for each voxel at the whole brain level.

4.1. Test-retest reliability results

The test-retest reliability analysis specifically aimed to capture possible repetition suppression effects due to stimuli repetition; for this reason, the subjects were scanned twice during the same day using the same stimuli (i.e., in the same parallel form) in both sessions, exacerbating possible habituation effects. When assessed with the voxel-wise test-retest reliability analysis, no statistically significant differences (pFWE < 0.05), neither in the positive nor in the negative condition, between the two sessions was observed.

Interestingly, the ROI-based analysis revealed a different reliability for positive and negative valence stimuli. Indeed, while for the positive conditions the ICC values ranged from negative ones (uninterpretable/approximately zero) (Berboth et al., 2021) to poor reliability, for the negative condition reliability was good to excellent in most of the ROIs considered. Moreover, the signal was highly reliable in both cortical areas belonging to frontal, occipital temporal and parietal cortices, and subcortical regions such as the hippocampus, the thalamus, and the amygdala. Our data suggest that signal reliability is independent from the considered ROI and anatomical areas but is instead highly dependent on the task condition (i.e., type of stimuli).

Table 3
Task elicited whole brain activity for the negative condition. The activation clusters are reported considering the cluster size, the statistics in terms of both t-value and p-value, the peak of activation and their localization according to the AAL atlas. Legend: FWE= family wise error correction fo multiple comparisons; equivk= cluster size. *R*= Right; *L*= Left; C= cortex.

p(FWE-corr)	equivk	T	x,y,z [mm]	AAL Label
<0.001	28,934	13.73	48 −68 10	R Middle Temporal C
		11.44	40 −42 −20	R Fusiform
		10.35	−38 −68 −14	L Fusiform
		11.37	24 −8 −16	R Hippocampus
		5.42	21 −14 −20	R Parahippocampus
		8.16	−25 −15 −15	L Hippocampus
		9.13	−30 −4 −20	L Amygdala
		6.43	12 −90 14	R Cuneus
		12.11	24 −96 −2	R Inferior Occipital C
		12.31	−46 −72 16	L Middle Occipital C
		8.91	−42 −76 −4	L Inferior Occipital C
		12.45	- 46 −76 4	R Middle Occipital C
		11.42	18 −98 6	R Superior Occipital C
		12.75	−40 −48 −18	L Inferior Temporal C
		11.02	42 −54 −14	R Inferior Temporal C
<0.001	2184	10.65	52 20 28	R Inferior Frontal Opercular C
		10.26	46 18 24	R Inferior Frontal Triangular C
		6.92	46 4 52	R Middle Frontal C
		6.57	52 10 44	R Precentral C
<0.001	912	9.26	−40 10 30	L Precentral C
		7.61	−42 22 18	L Inferior Frontal Triangular C
		7.31	−44 12 27	L Inferior Frontal Opercular C
<0.001	941	8.41	−6 54 42	L Superior Medial Frontal C
		6.44	4 52 32	R Superior Medial Frontal C
		5.29	10 26 64	R Supplementary Motor Area
		7.18	−10 56 34	L Superior Frontal C
<0.001	469	7.02	6 −8 6	R Thalamus
		6.78	−4 −12 6	L Thalamus
		6.21	14 −2 14	R Caudate
		5.93	−10 2 14	L Caudate
<0.001	138	6.9	28 −48 50	R Inferior Parietal
<0.001	61	5.85	−52 −4 −16	L Postcentral C
		5.69	−52 −16 −12	L Middle Temporal C
0.002	35	5.61	−6 −48 44	L Precuneus

Our results are novel and differ from previous studies investigating emotion stimulation tasks. Mc Dermott et al. (McDermott et al., 2020) reported greater reliability for visual areas, which are involved in the sensory processing of the stimuli, compared to areas designated to emotional processing; in this case, the differences with respect to our results, could be ascribed to the different task paradigm employed by McDermott and colleagues which investigated emotions through a face processing task, implying a more coherent visual perceptual load of the pictures (i.e., only faces) compared to the stimuli used here, depicting different objects and scenarios. Similarly, Berboth et al. (Berboth et al., 2021), in an emotion regulation task, found higher reliability of the signal in cortical, compared to subcortical, regions, in line with previous findings of a recent meta-analysis (Elliott et al., 2020). However, the ICC values reported in (Berboth et al., 2021) are in line with the one obtained in our study, especially when considering that they employed equivalent stimuli, in order to mitigate possible habituation effects, as we also did.

Of note, the classification of the signal reliability level used in the present study was derived according to (Cicchetti and Sparrow, 1981). This classification, however, has been defined for behavioural and psychological measures in which ICC is considered reliable when equal or higher to 0.8. Such values are not often achievable for fMRI measures (Hedge et al., 2018).

Table 4
The coordinates of the AAL atlas labels for the selected ROI are reported in the table for the positive and the negative contrasts. A total of 18 ROIs has been selected for the positive contrast while 22 ROIs were derived from the negative one. Legend: ROI= regions of interest; *P*= Positive; *N*= Negative; *R*= Right; *L*= Left; C= Cortex.

Lobe	AAL Labels	Positive Contrast ROI Number	Negative Contrast ROI Number
Frontal Cortex	L Inferior Frontal Opercular C	P1	N1
	L Inferior Frontal Triangular C	P2	N2
	L Orbitofrontal C	P3	–
	L Superior Medial Frontal C	–	N3
	R Superior Medial Frontal C	–	N4
	R Inferior Frontal Opercular C	P4	N5
	R Inferior Frontal Triangular C	P5	N6
Insular Cortex	L insula	P6	–
Occipital Cortex	L Inferior Occipital C	P7	N7
	L Middle Occipital C	P8	N8
	R Inferior Occipital C	P9	N9
	R Middle Occipital C	P10	N10
	R Cuneus	P11	N11
Parietal Cortex	L Precuneus	–	N12
	R Inferior Parietal	–	N13
Temporal Cortex	L Fusiform	P12	N14
	L Middle Temporal C	P13	N15
	L Hippocampus	–	N16
	R Fusiform	P14	–
	R Middle Temporal C	P15	N17
	R Hippocampus	P16	N18
	R Parahippocampus	–	N19
Subcortical	L Amygdala	–	N20
	L Thalamus	P17	N21
	R Thalamus	P18	N22

4.2. Reproducibility results

The reproducibility analysis was essentially aimed to assess and validating a possible strategy to minimize repetition suppression effects in fMRI longitudinal studies employing visual stimuli with different emotional valence. Thus, the subjects were scanned twice with a mean inter-scan interval of 4 months and using two parallel forms of the same task. The time interval was chosen according to the average duration of a typical rehabilitation treatment.

This analysis revealed no differences at the voxel-wise level. However, when considering the ROI-based analysis the ICC values showed heterogeneous results both within and between conditions (i.e., emotional valence).

Overall, for the reproducibility analysis, a higher degree of signal variability relative to the test-retest condition was detected in both positive and negative conditions. Specifically, the ICC values for the positive condition were overall higher compared to the ones obtained considering test-retest reliability, suggesting that the differences may be partially ascribed to a neural suppression and habituation phenomenon which was at least partially mitigated in the reproducibility analysis by the employment of different equivalent stimuli of the two parallel forms and the elapsed time. Conversely, the negative stimuli showed more heterogeneous ICC values with respect to test-retest analysis but still preserving on average higher reliability compared to the positive condition.

Altogether, these results suggest that the prevailing mechanism affecting within-subject signal reliability is related to task condition suggesting the employment of diverse neural strategies and circuits over time. Evidence on the role of neural strategies in fMRI test-retest signal reliability, was recently investigated by (Berboth et al., 2021) during a negative emotion regulation task. The authors reported a decrease in neural activity across sessions at both whole-brain and region-wise level. Since equivalent stimuli were used between the different sessions, preventing the habituation effect, the authors ascribed the observed results to a reduced cognitive effort and, in turn, to an enhanced neural efficiency in emotion regulatory processes.

Furthermore, only for the positive condition, the higher reproducibility retrieved when using two parallel forms of the task and after a wide time interval, suggests that a small component of fMRI adaptation could still be present. However, the improvement in positive condition reliability could be also due to different sample sizes and therefore further studies are needed to confirm this aspect.

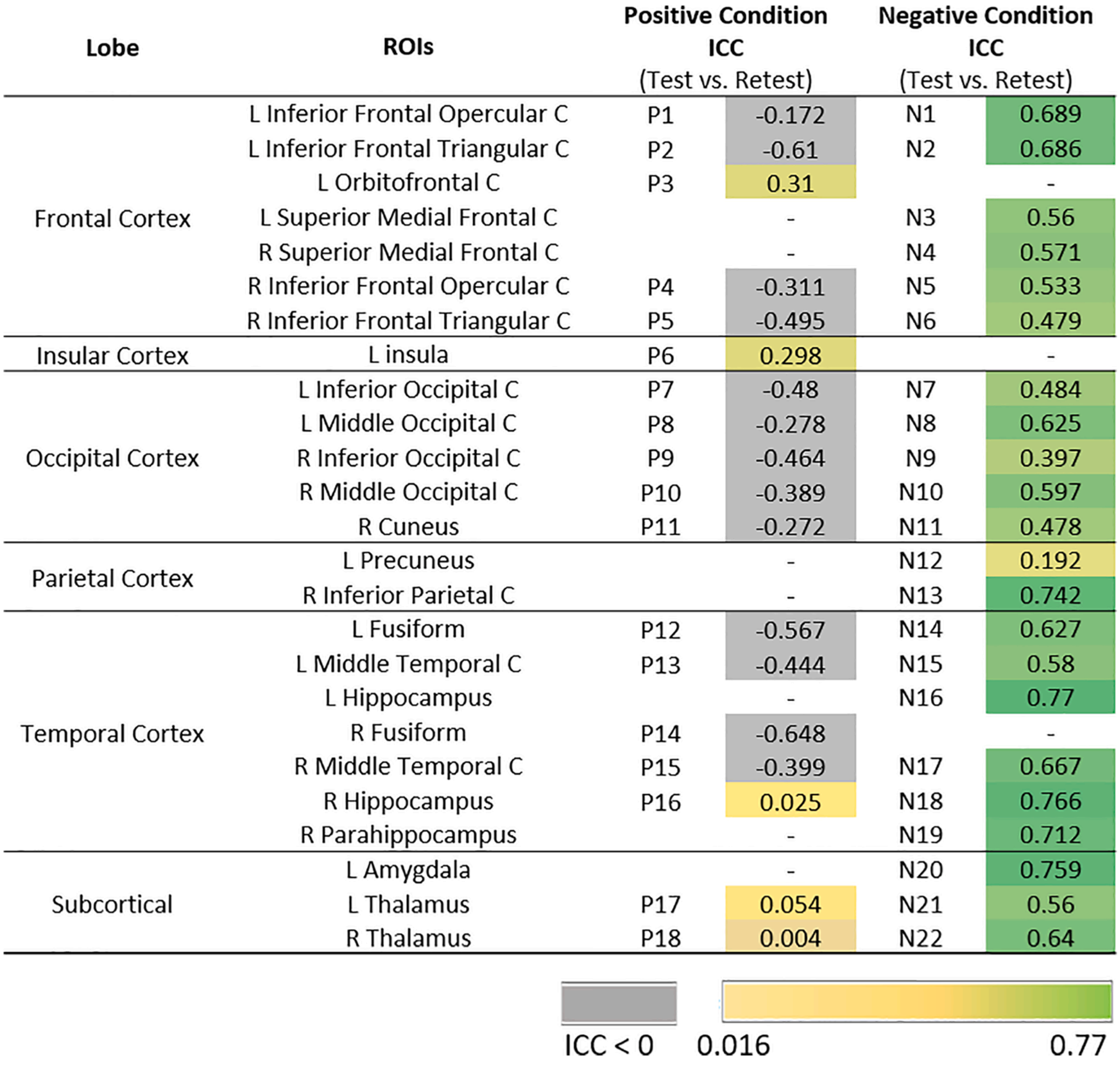
Notably, our results also highlight different degrees of reliability and reproducibility levels between the two experimental conditions (i.e., positive or negative emotional stimuli). These results are in line with previous imaging studies revealing that fMRI signal variability depends upon the emotional valence of the proposed stimuli (Dores et al., 2013; Mourão-Miranda et al., 2003). Specifically, the observed higher reliability of fMRI signal when using negative stimuli could be related to the well-known '*unpleasant emotion bias*' (Dores et al., 2013). In fact, it is well established that unpleasant/negative stimuli produce higher neural activation with respect to neutral and pleasant/positive ones. It has also been demonstrated that higher levels of activation may result in higher ICC values, thus explaining the differences we observed depending on the experimental conditions (i.e., emotional valence) (Berboth et al., 2021; Caceres et al., 2009; Fliessbach et al., 2010; Korucuoglu et al., 2021).

The new evidence reported here corroborates the hypothesis that when assessing the reliability of emotion stimulation tasks, the most impacting factor to consider is related to the individual (within-subject) variability due to the employed strategy in coping with the administered stimuli more than to repetition suppression and habituation phenomena.

Furthermore, our results suggest that when implementing emotion fMRI tasks to monitor both a neurological/neuropsychological pathology and/or a pharmacological or rehabilitative treatment, negative stimuli might be more effective in capturing differences among clinical populations or between treatment time points. However, in a life-span perspective future studies including subjects with broader age span and clinical conditions are needed to ensure the generalizability of these results.

Finally, in this work we conducted a reliability analysis using both GLM and ICC analyses. The different (complementary) role of GLM and ICC analyses in the context of the conducted reliability analysis allowed a more complete picture of the problem but it is essential to point out these roles. In the GLM analysis, the interpretation differs depending on the hypothesis made on the (regional) effects. Assuming that no difference should exist in the fMRI signal from a stimulus-activated region between two repeated sessions, any significant regional effect should be interpreted as an indication of poor reproducibility. This might be either due to the technical (e.g., noise, artefacts) or neurophysiological effects (e.g., habituation, anticipation), thereby the presence of a significant regional effect would warn against targeting that region to probe any interventional effects in the case of treatment. However, even in the absence of significant regional effects in the GLM analysis, as shown in this study, a complete reliability statement or picture would still require an ICC analysis to consider the occurrence of both significant and non-significant regional effects.

Future works are warranted to comprehensively address the impact of coping strategies on fMRI signal reliability also considering the emotional valence of the stimuli (i.e., positive vs. negative). One possible limitation of the study is that we used an implicit emotional stimulation of the subjects, i.e., without asking the participants for an explicit rating of the stimuli on-line nor debriefing them off-line to verify how they would have rated each picture. However, the used stimuli were chosen among the extreme values of the normative ranges for each emotional valence to minimise the possibility that subjects would



Lobe	ROIs	Positive Condition ICC (Test vs. Retest)		Negative Condition ICC (Test vs. Retest)	
Frontal Cortex	L Inferior Frontal Opercular C	P1	-0.172	N1	0.689
	L Inferior Frontal Triangular C	P2	-0.61	N2	0.686
	L Orbitofrontal C	P3	0.31		-
	L Superior Medial Frontal C		-	N3	0.56
	R Superior Medial Frontal C		-	N4	0.571
	R Inferior Frontal Opercular C	P4	-0.311	N5	0.533
	R Inferior Frontal Triangular C	P5	-0.495	N6	0.479
Insular Cortex	L insula	P6	0.298		-
Occipital Cortex	L Inferior Occipital C	P7	-0.48	N7	0.484
	L Middle Occipital C	P8	-0.278	N8	0.625
	R Inferior Occipital C	P9	-0.464	N9	0.397
	R Middle Occipital C	P10	-0.389	N10	0.597
	R Cuneus	P11	-0.272	N11	0.478
Parietal Cortex	L Precuneus		-	N12	0.192
	R Inferior Parietal C		-	N13	0.742
Temporal Cortex	L Fusiform	P12	-0.567	N14	0.627
	L Middle Temporal C	P13	-0.444	N15	0.58
	L Hippocampus		-	N16	0.77
	R Fusiform	P14	-0.648		-
	R Middle Temporal C	P15	-0.399	N17	0.667
	R Hippocampus	P16	0.025	N18	0.766
	R Parahippocampus		-	N19	0.712
Subcortical	L Amygdala		-	N20	0.759
	L Thalamus	P17	0.054	N21	0.56
	R Thalamus	P18	0.004	N22	0.64

Fig. 4. The interclass correlation coefficient (ICC) derived from the test-retest ROI-based analysis is reported in the figure for the positive and negative conditions. Negative (uninterpretable) ICC values are reported in gray. Positive ICC values are color-coded from yellow (lowest ICC) to green (highest ICC). **Legend:** ICC= intraclass correlation coefficient; ROI= regions of interest; *R*= Right; *L*= Left; C= Cortex, *P*= Positive, *N*= Negative.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

intrinsically rate each picture differently from what was to be expected and to ensure proper emotional stimulation at selected valence.

In particular, the subjective emotional valence could also have a role on the different coping strategies. Another important aspect to be considered, would be the intra-scan fMRI signal reliability. Future studies are needed to precisely investigate also this aspect which was out of the scope of the present work mainly focusing on the fMRI signal reliability specifically referred to longitudinal settings. Finally, in analogy to what observed during fMRI studies involving cognitive (Cabinio et al., 2015; Castelli et al., 2010; Farina et al., 2017) and motor tasks (Di Tella et al., 2021) which showed age-related differences in the activation patterns, it will be important to extend the study of fMRI signal reliability to different age groups, for instance during development and aging populations.

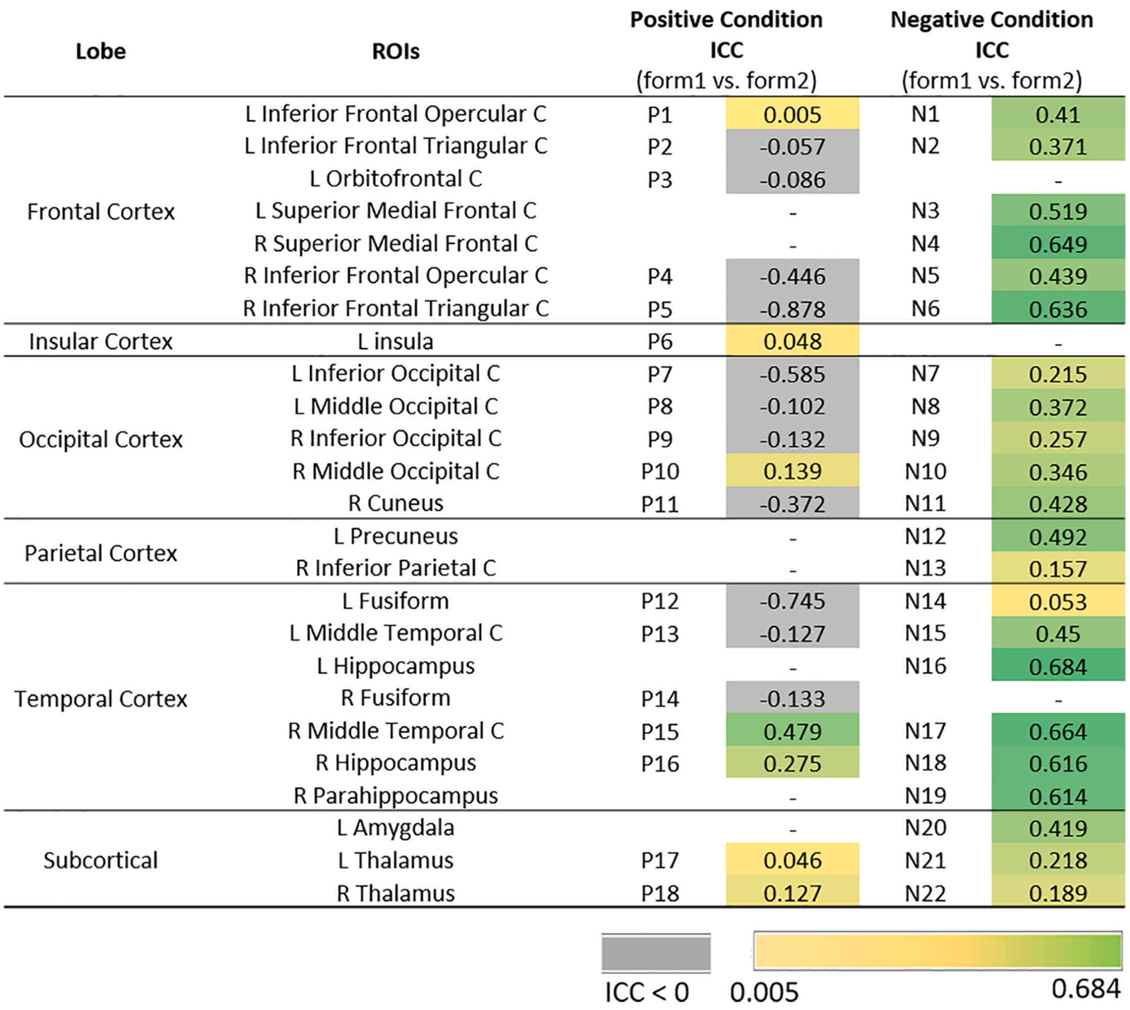
5. Conclusion

Overall, our study investigated the relative contribution of two concurrent mechanisms affecting the signal reliability of task-fMRI measures. The implementation of two parallel forms allowed us to exclude predominant effects of the so-called repetition suppression or habituation phenomena impacting the reliability of the fMRI signal. Thus, the main effect responsible for low signal reliability seems to be ascribed to the different neural strategy involved when repeating the task. This might be related with coping strategies aimed to reduce the cognitive/emotional effort elicited by complex tasks. It is worth mentioning, that the voxel-wised approach with GLM analysis failed to capture the poor signal reliability which was instead highlighted using the ROI-based ICC analysis. The use of the GLM analysis in longitudinal studies in the absence of a previous reliability assessment could yield to misleading (false positive or false negative) results, on the effects of a given treatment or of disease progression.

To conclude, our study demonstrated the importance of investigating the signal reliability of fMRI tasks before implementing longitudinal paradigms.

CRediT authorship contribution statement

Alice Pirastru: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Sonia Di Tella:** Methodology, Writing – review & editing. **Marta Cazzoli:** Data curation, Writing – review & editing. **Fabrizio Esposito:** Writing – review & editing, Supervision. **Giuseppe Baselli:** Writing – review & editing, Supervision. **Francesca Baglio:** Conceptualization, Writing – review &



Lobe	ROIs	Positive Condition ICC (form1 vs. form2)		Negative Condition ICC (form1 vs. form2)	
Frontal Cortex	L Inferior Frontal Opercular C	P1	0.005	N1	0.41
	L Inferior Frontal Triangular C	P2	-0.057	N2	0.371
	L Orbitofrontal C	P3	-0.086		-
	L Superior Medial Frontal C		-	N3	0.519
	R Superior Medial Frontal C		-	N4	0.649
	R Inferior Frontal Opercular C	P4	-0.446	N5	0.439
	R Inferior Frontal Triangular C	P5	-0.878	N6	0.636
Insular Cortex	L insula	P6	0.048		-
Occipital Cortex	L Inferior Occipital C	P7	-0.585	N7	0.215
	L Middle Occipital C	P8	-0.102	N8	0.372
	R Inferior Occipital C	P9	-0.132	N9	0.257
	R Middle Occipital C	P10	0.139	N10	0.346
	R Cuneus	P11	-0.372	N11	0.428
Parietal Cortex	L Precuneus		-	N12	0.492
	R Inferior Parietal C		-	N13	0.157
Temporal Cortex	L Fusiform	P12	-0.745	N14	0.053
	L Middle Temporal C	P13	-0.127	N15	0.45
	L Hippocampus		-	N16	0.684
	R Fusiform	P14	-0.133		-
	R Middle Temporal C	P15	0.479	N17	0.664
	R Hippocampus	P16	0.275	N18	0.616
	R Parahippocampus		-	N19	0.614
Subcortical	L Amygdala		-	N20	0.419
	L Thalamus	P17	0.046	N21	0.218
	R Thalamus	P18	0.127	N22	0.189

Fig. 5. The interclass correlation coefficient (ICC) derived from the reproducibility ROI-based analysis is reported in the figure for the positive and negative conditions. Negative (uninterpretable) ICC values are reported in gray. Positive ICC values are color-coded from yellow (lowest ICC) to green (highest ICC). **Legend:** ICC= intraclass correlation coefficient; ROI= regions of interest; *R*= Right; *L*= Left; C= Cortex, *P*= Positive, *N*= Negative.

editing, Supervision, Funding acquisition. **Valeria Blasi:** Conceptualization, Supervision, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data availability

Data will be made available on request.

Funding and Acknowledgements

Work supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) – A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120457.

References

Anzuino, I., Baglio, F., Pelizzari, L., Cabinio, M., Biassoni, F., Gnerre, M., Di Tella, S., 2023. Production of emotions conveyed by voice in Parkinson's disease: Association between variability of fundamental frequency and gray matter volumes of regions involved in emotional prosody. Neuropsychology. https://doi.org/10.1037/neu0000912.

Bartko, J.J., 1976. On various intraclass correlation reliability coefficients. Psychol. Bull. 83 (5), 762.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N.Y. Acad. Sci. 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Berboth, S., Windischberger, C., Kohn, N., Morawetz, C., 2021. Test-retest reliability of emotion regulation networks using fMRI at ultra-high magnetic field. Neuroimage 232, 117917. https://doi.org/10.1016/j.neuroimage.2021.117917.

Blonder, L.X., Slevin, J.T., 2011. Emotional dysfunction in Parkinson's disease. Behav. Neurol. 24 (3), 201–217. https://doi.org/10.3233/ben-2011-0329.

Bora, E., Velakoulis, D., Walterfang, M., 2016. Meta-Analysis of Facial Emotion Recognition in Behavioral Variant Frontotemporal Dementia: Comparison With Alzheimer Disease and Healthy Controls. J. Geriatr. Psychiatry Neurol. 29 (4), 205–211. https://doi.org/10.1177/0891988716640375.

Bora, E., Yener, G.G., 2017. Meta-Analysis of Social Cognition in Mild Cognitive Impairment. J. Geriatr. Psychiatry Neurol. 30 (4), 206–213. https://doi.org/10.1177/0891988717710337.

Cabinio, M., Rossetto, F., Blasi, V., Savazzi, F., Castelli, I., Massaro, D., Baglio, F., 2015. Mind-Reading Ability and Structural Connectivity Changes in Aging. Front Psychol. 6, 1808. https://doi.org/10.3389/fpsyg.2015.01808.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Castelli, I., Baglio, F., Blasi, V., Alberoni, M., Falini, A., Liverta-Sempio, O., Marchetti, A., 2010. Effects of aging on mindreading ability through the eyes: an fMRI study. Neuropsychologia 48 (9), 2586–2594. https://doi.org/10.1016/j.neuropsychologia.2010.05.005.

Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Defic. 86 (2), 127–137.

Di Tella, S., Blasi, V., Cabinio, M., Bergsland, N., Buccino, G., Baglio, F., 2021. How Do We Motorically Resonate in Aging? A Compensatory Role of Prefrontal Cortex. Front Aging. Neurosci. 13, 694676 https://doi.org/10.3389/fnagi.2021.694676.

Dores, A.R., Almeida, I., Barbosa, F., Castelo-Branco, M., Monteiro, L., Reis, M., Caldas, A.C., 2013. Effects of emotional valence and three-dimensionality of visual stimuli on brain activation: an fMRI study. NeuroRehabilitation 33 (4), 505–512. https://doi.org/10.3233/nre-130987.

Drobyshevsky, A., Baumann, S.B., Schneider, W., 2006. A rapid fMRI task battery for mapping of visual, motor, cognitive, and emotional function. Neuroimage 31 (2), 732–744. https://doi.org/10.1016/j.neuroimage.2005.12.016.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Hariri, A. R., 2020. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. Psychol. Sci. 31 (7), 792–806. https://doi.org/10.1177/0956797620916786.

Farina, E., Baglio, F., Pomati, S., D'Amico, A., Campini, I.C., Di Tella, S., Pozzo, T., 2017. The Mirror Neurons Network in Aging, Mild Cognitive Impairment, and Alzheimer Disease: A functional MRI Study. Front Aging. Neurosci. 9, 371. https://doi.org/10.3389/fnagi.2017.00371.

Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50 (3), 1168–1176. https://doi.org/10.1016/j.neuroimage.2010.01.036.

Fostering reproducible fMRI research. Nat. Commun. 8, 2017, 14748. https://doi.org/10.1038/ncomms14748.

Giraudeau, B., 1996. Negative values of the intraclass correlation coefficient are not theoretically possible. J. Clin. Epid.emiology 49 (10), 1205.

Gray, H.M., Tickle-Degnen, L., 2010. A meta-analysis of performance on emotion recognition tasks in Parkinson's disease. Neuropsychology 24 (2), 176–191. https://doi.org/10.1037/a0018104.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10 (1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Meth. 50 (3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

Heilicher, M., Crombie, K.M., Cisler, J.M., 2022. Test-retest reliability of fMRI during an emotion processing task: Investigating the impact of analytical approaches on ICC values. Front Neuroimaging 1. https://doi.org/10.3389/fnimg.2022.859792.

Holiga, Š., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R.J., Dukart, J., 2018. Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. PLoS One 13 (11), e0206583. https://doi.org/10.1371/journal.pone.0206583.

Jenkinson, M., Pechaud, M., Smith, S., 2005. BET2: MR-based estimation of brain, skull and scalp surfaces. In: Eleventh annual meeting of the organization for human brain mapping.

Klein-Koerkamp, Y., Beaudoin, M., Baciu, M., Hot, P., 2012. Emotional decoding abilities in Alzheimer's disease: a meta-analysis. J. Alzheimers Dis. 32 (1), 109–125. https://doi.org/10.3233/jad-2012-120553.

Korucuoglu, O., Harms, M.P., Astafiev, S.V., Golosheykin, S., Kennedy, J.T., Barch, D.M., Anokhin, A.P., 2021. Test-Retest Reliability of Neural Correlates of Response Inhibition and Error Monitoring: An fMRI Study of a Stop-Signal Task. Front Neurosci. 15, 624911 https://doi.org/10.3389/fnins.2021.624911.

Lahey, M.A., Downey, R.G., Saal, F.E., 1983. Intraclass correlations: There's more there than meets the eye. Psychol. Bull. 93 (3), 586.

Laird, A.R., Eickhoff, S.B., Kurth, F., Fox, P.M., Uecker, A.M., Turner, J.A., Fox, P.T., 2009. ALE Meta-Analysis Workflows Via the Brainmap Database: Progress Towards A Probabilistic Functional Brain Atlas. Front Neuroinform. 3, 23. https://doi.org/10.3389/neuro.11.023.2009.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

McDermott, T.J., Kirlic, N., Akeman, E., Touthang, J., Cosgrove, K.T., DeVille, D.C., Aupperle, R.L., 2020. Visual cortical regions show sufficient test-retest reliability while salience regions are unreliable during emotional face processing. Neuroimage 220, 117077. https://doi.org/10.1016/j.neuroimage.2020.117077.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Meth. 1, 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

Morawetz, C., Riedel, M.C., Salo, T., Berboth, S., Eickhoff, S.B., Laird, A.R., Kohn, N., 2020. Multiple large-scale neural networks underlying emotion regulation. Neurosci. Biobehav. Rev. 116, 382–395. https://doi.org/10.1016/j.neubiorev.2020.07.001.

Mourão-Miranda, J., Volchan, E., Moll, J., de Oliveira-Souza, R., Oliveira, L., Bramati, I., Pessoa, L., 2003. Contributions of stimulus valence and arousal to visual activation during emotional perception. Neuroimage 20 (4), 1955–1963. https://doi.org/10.1016/j.neuroimage.2003.08.011.

Noble, S., Scheinost, D., Constable, R.T., 2021. A guide to the measurement and interpretation of fMRI test-retest reliability. Curr. Opin. Behav. Sci. 40, 27–32. https://doi.org/10.1016/j.cobeha.2020.12.012.

Pfaff, L., Lamy, J., Noblet, V., Gounot, D., Chanson, J.B., de Seze, J., Blanc, F., 2019. Emotional disturbances in multiple sclerosis: A neuropsychological and fMRI study. Cortex 117, 205–216. https://doi.org/10.1016/j.cortex.2019.02.017.

Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M., 2020. Automated anatomical labelling atlas 3. Neuroimage 206, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189.

Segaert, K., Weber, K., de Lange, F.P., Petersson, K.M., Hagoort, P., 2013. The suppression of repetition enhancement: a review of fMRI studies. Neuropsychologia 51 (1), 59–66. https://doi.org/10.1016/j.neuropsychologia.2012.11.006.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155. https://doi.org/10.1002/hbm.10062.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imag. 29 (6), 1310–1320. https://doi.org/10.1109/tmi.2010.2046908.

Xia, M., Wang, J., He, Y., 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One 8 (7), e68910. https://doi.org/10.1371/journal.pone.0068910.