Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Test–retest reliability of the prefrontal response to affective pictures based on functional near-infrared spectroscopy

Yuxia Huang
Mengchai Mao
Zong Zhang
Hui Zhou
Yang Zhao
Lian Duan
Ute Kreplin
Xiang Xiao
Chaozhe Zhu

Yuxia Huang, Mengchai Mao, Zong Zhang, Hui Zhou, Yang Zhao, Lian Duan, Ute Kreplin, Xiang Xiao, Chaozhe Zhu, "Test–retest reliability of the prefrontal response to affective pictures based on functional near-infrared spectroscopy," *J. Biomed. Opt.* **22**(1), 016011 (2017), doi: 10.1117/1.JBO.22.1.016011.

SPIE.

Test–retest reliability of the prefrontal response to affective pictures based on functional near-infrared spectroscopy

Yuxia Huang,[a] Mengchai Mao,[a,†] Zong Zhang,[a] Hui Zhou,[a] Yang Zhao,[a] Lian Duan,[a] Ute Kreplin,[b] Xiang Xiao,[a] and Chaozhe Zhu[a,*]

[a]Beijing Normal University, State Key Laboratory of Cognitive Neuroscience and Learning, 19 Xin Jie Kou Wai Da Jie, Hai Dian District, Beijing 100875, China

[b]Massey University, School of Psychology, 3.26 Psychology Building, Tennent Drive, Palmerston North 4474, Manawalu, New Zealand

Abstract. Functional near-infrared spectroscopy (fNIRS) is being increasingly applied to affective and social neuroscience research; however, the reliability of this method is still unclear. This study aimed to evaluate the test–retest reliability of the fNIRS-based prefrontal response to emotional stimuli. Twenty-six participants viewed unpleasant and neutral pictures, and were simultaneously scanned by fNIRS in two sessions three weeks apart. The reproducibility of the prefrontal activation map was evaluated at three spatial scales (mapwise, clusterwise, and channelwise) at both the group and individual levels. The influence of the time interval was also explored and comparisons were made between longer (intersession) and shorter (intrasession) time intervals. The reliabilities of the activation map at the group level for the mapwise (up to 0.88, the highest value appeared in the intersession assessment) and clusterwise scales (up to 0.91, the highest appeared in the intrasession assessment) were acceptable, indicating that fNIRS may be a reliable tool for emotion studies, especially for a group analysis and under larger spatial scales. However, it should be noted that the individual-level and the channelwise fNIRS prefrontal responses were not sufficiently stable. Future studies should investigate which factors influence reliability, as well as the validity of fNIRS used in emotion studies. [DOI: 10.1117/1.JBO.22.1.016011]

Keywords: test–retest; reproducibility; near-infrared spectroscopy; emotion; prefrontal cortex.

Paper 160528RR received Aug. 4, 2016; accepted for publication Dec. 27, 2016; published online Jan. 17, 2017.

1 Introduction

Functional near-infrared spectroscopy (fNIRS) is a new non-invasive neuroimaging technique that has attracted increasing attention in recent years. It shines near-infrared light into the outer cerebral cortex and the light absorption rate varies with the change of hemoglobin intensity.[1] Thus, a certain degree of inference of the neural and psychological activities that lead to a change in hemoglobin intensity can be deduced based on the measurement of the change in light intensity. FNIRS has an acceptable spatial and temporal resolution for current neuroimaging studies. Furthermore, its good portability and small bondage of body movements make it possible to carry out experiments in near-natural circumstances that enhance the ecological validity of studies, especially those exploring affective and social issues. The ability to use "friendly" experimental environments also provides more convenience for special populations when performing neuroimaging studies, such as infants,[2,3] children,[4,5] pregnant women,[6] patients with psychiatric disorders,[7–9] and animals.[10,11]

Hoshi and Chen[4] used fNIRS to observe regional cerebral blood flow changes in children with different emotions. This study introduced fNIRS to the field of affective neuroscience for the first time. Since then, there have been considerable numbers of emotion studies conducted using the fNIRS technique. For example, some studies[12–15] observed cerebral responses to emotional stimuli, such as affective pictures, emotional faces, and affective sounds. Some studies[16–18] examined the association between mood states and cognition. In studies on emotion regulation, activation of the prefrontal cortex was reported.[5,19] Studies of interpersonal affective interactions[2,20,21] have also benefited from the near-natural data acquisition environment of fNIRS. This technique was also used in studies on emotion decoding[22–24] and enhanced the future practicability of these kinds of studies.

Taken together, the literature indicates that fNIRS may be a promising tool in the field of affective and social neuroscience. However, although there has been increasing interest and application of fNIRS in the affective and soical scope, there has not been a report of fNIRS's reliability in emotion studies. Because it is a new technique, its reliability is the foundation of further extensive applications. When researchers evaluate the advantages and disadvantages of a measurement tool, its reliability is always an essential aspect. Some of the emotion studies using fNIRS have reported similar conclusions in some research topics. However, it is hard to evaluate the fNIRS reliability strictly and directly based on these similarities, because these studies may have great differences with regards to the participants, stimulation materials, experimental paradigms, data acquisition and analysis methods, and experimental equipment. There have been several reports of fNIRS reliability when used to measure psychological functions, such as visual,[25] motor,[26,27] and executive[28,29] functions, and the network features of the

*Address all correspondence to: Chaozhe Zhu, E-mail: czzhu@bnu.edu.cn

†Joint first author.

1083-3668/2017/$25.00 © 2017 SPIE

human brain.[30,31] In these reports, both the fNIRS device and the data acquisition and analysis methods had acceptable test–retest reliability. But these results cannot be directly extrapolated to the field of emotion research because emotion studies have significant differences to the previously mentioned studies in terms of the research questions, experimental tasks, and brain areas studied. In addition, emotion phenomena are very subtle and complex, which has the salient properties of a black box. In this situation, it is necessary to conduct a pointed study to investigate the reliability of fNIRS in emotion research. FNIRS, similar to functional magnetic resonance imaging (fMRI), is a blood oxygenation level dependent (BOLD) neuroimaging technique. Currently, there have been a few reliability reports of fMRI used in measurements of emotion-related BOLD signals. For example, some studies[32–34] indicated relatively stable amygdala BOLD responses to emotional faces at the group level but lower reliability at the subjective-specific level. There was also a report of robust and reliable emotion-dependent amygdala habituation by fMRI measurement.[35] Good reliability was also reported for the node degree of limbic nodes by graph theory analyses in an emotional face-matching task.[36] In these studies, the brain area typically concerned is the amygdala. However, current fNIRS can only measure the cortical surface and cannot detect the subcortical areas. In addition, there is a difference in the imaging principle between fNIRS and fMRI. Thus, it is hard to deduce fNIRS reliability in emotion studies based on previous fMRI reports.

Currently, fNIRS has been used in the research field of affective and social neuroscience. Its reproducibility is one of the prerequisites for further applications, such as the search for emotional neuromarkers and the quantitative monitoring of mental status. Thus, the present study aimed to test the reliability of fNIRS when applied to an emotion task. In emotional research, the neural response to emotional stimuli is an important issue in the field of affective neuroscience. Currently, this issue has also been investigated by a large number of emotion studies using fNIRS, among which the function of the prefrontal cortex is most frequently concerned.[8,12,14,15,37–40] Therefore, the current study focused on evaluating the reliability of prefrontal activation in response to affective pictures,[41] which are widely used to elicit emotions in laboratories. Previous studies have paid more attention to negative (i.e., unpleasant) emotions; therefore, the current study followed this tradition and evaluated the reliability of the prefrontal responses elicited by unpleasant pictures with a neutral condition as a contrast. The fNIRS reproducibility of emotion-related prefrontal activation was evaluated by a test–retest reliability frequently used in reliability assessments. The time interval between test and retest reported in previous studies varied from a few minutes,[30] to a few weeks[31,33] and a few months,[26,42] and even over 1 year.[43] The current study set three weeks, a temporal length in the middle of previously used timepoints as the test–retest interval. For a comprehensive assessment, the current study evaluated the fNIRS reproducibility from global to localized scales at both the group and individual levels.

2 Materials and Methods

2.1 *Participants*

Twenty-eight university students participated as paid volunteers. Data from two participants were excluded from further processing due to participant loss and data-recording failure, respectively. Thus, data from 26 participants (11 females and 15 males, mean age 22.4 ± 2.1 years) were included in subsequent analyses. Participants were scanned in two sessions three weeks (mean interval 21.0 ± 0.9 days) apart. All participants were right-handed, as assessed using the Edinburgh Handedness Inventory,[44] and had normal or corrected-to-normal vision. They reported no history of neurological or mental health problems. Written informed consent was obtained from all of the participants. This study was approved by the research ethics committee of the School of Brain and Cognitive Sciences in Beijing Normal University. All study procedures were conducted in accordance with the latest version of the Declaration of Helsinki.

2.2 *Materials*

Emotion-evocative pictures: 32 unpleasant (mean valence 2.44 ± 0.68, mean arousal 5.85 ± 0.64) and 32 neutral pictures (mean valence 5.03 ± 0.31, mean arousal 2.89 ± 0.66) were selected from the International Affective Picture System[41] for use during this study.

Positive and Negative Affect Schedule (PANAS): the PANAS scale [45,46] was used to measure participants' affective states before the experiment.

Self-Assessment Manikin (SAM): the nine-point SAM system[41,47] was used to measure participants' emotional experiences (the valence dimension and the arousal dimension) elicited by the affective pictures.

2.3 *Procedure*

In the first experimental session, on attending the laboratory, the experimenter briefly introduced participants to the experiment. Then participants were asked to complete the PANAS before beginning the main part of the study, during which fNIRS scanning was conducted. Then the experimenter explained the main task to participants in detail. The meaning of valence and arousal, and how to use SAM to make self-reports of subjective feelings were explained. The formal experiment began when participants understood the task and familiarized themselves with the reporting operation through practice. The scanning was carried out in a quiet room with dim light. Participants were seated comfortably in front of a 19-in. monitor. As shown in Fig. 1, their task began with a 30-s resting-state fNIRS scan with the mind relaxed and eyes open. Then a "+" fixation point was presented for 10 s to remind participants to pay attention to the upcoming pictures. Afterwards, eight blocks (four unpleasant blocks and four neutral blocks) of pictures were presented. The unpleasant block and the neutral block were presented alternatively and the same category of block would not appear in two consecutive times. Each block consisted of four pictures in the same valence category. Each picture was presented for 6 s thus, each block lasted 24 s. Participants were required to view the pictures carefully and try to immerse themselves in the scenes presented. After each block, they were asked to report their emotional experiences during the picture-viewing segment using the nine-point SAM scales. The SAM picture was being presented until participants had clicked the mouse to make a choice. After the self-reporting, there was a 24-s resting period with a "+" presented on the screen. Participants could take a short break after finishing the eight blocks (the first half of an experimental session). In the second half of the experiment, as for the first half, participants underwent a 30-s resting period

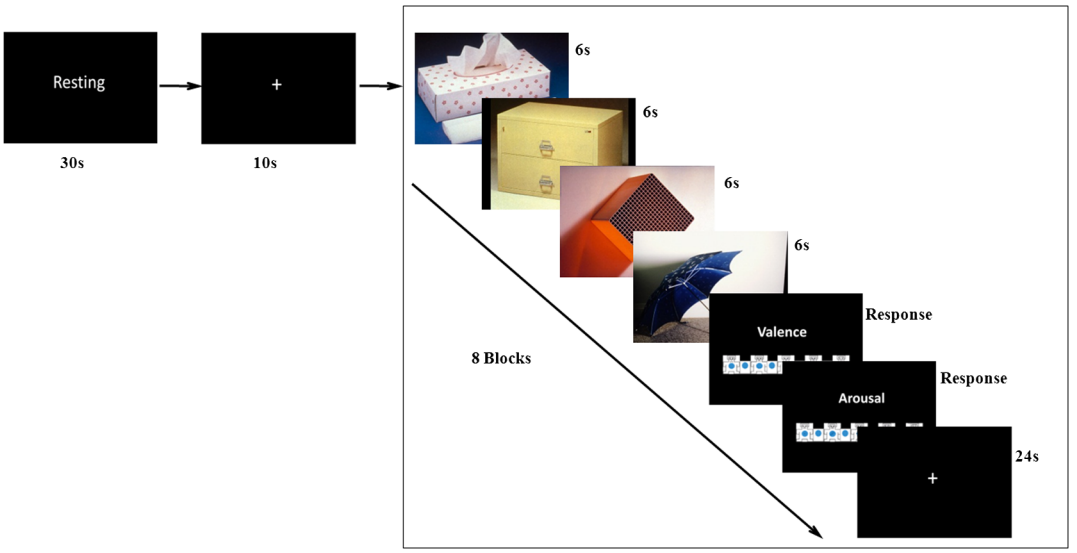

Fig. 1 A schematic diagram of the experimental processes. One experimental session consisted of the same two processes. One half of an experimental session consisted of eight blocks (four unpleasant and four neutral blocks) of picture viewing and self-reporting. Here, a neutral block was used as an example.

and a 10-s fixation period, then viewed the affective pictures and reported their feelings for a further eight blocks. The pictures were presented, and self-report ratings were collected, via E-Prime 2.0. Each affective picture would be presented only once. Participants underwent the same experimental procedures in the second session after a three-week interval.

2.4 *fNIRS Measurements*

The fNIRS measurements were conducted with a 57-channel fNIRS system (LABNIRS, Shimadzu Co., Japan) with 18 emitters and 18 detectors (interoptode distance = 30 mm) arrayed as illustrated in Fig. 2(b). A holder cap was used to place the optodes upon the frontal cortex, based on the international 10–20 system. The optode between channels 47 and 48 was placed at position FPz, and channels 32 and 42 were placed near positions T3 and T4, respectively. To maintain the locations of optodes and channels consistently through the two experimental sessions, the holder positioning for both sessions was completed by the same well-trained experimenter. To estimate the cortical locations for the corresponding channels, the scalp positions of each optode and each channel were recorded using a 3-D magnetic digitizer (PATRIOT, Polhemus Inc.) on one of the participants. Then a probabilistic registration process[48] was performed to estimate the cortical sites using NIRS_SPM 4.0.[49]

The absorption of near-infrared light at three wavelengths (780, 805, and 830 nm) was measured with a sampling rate of 17.54 Hz. Changes in the concentrations of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) were computed using the modified Beer–Lambert law. The total hemoglobin (HbT) signal was the sum of the HbO and HbR signals.

2.5 *fNIRS Data Preprocessing and Analysis*

The fNIRS data derived from each participant included four segments: data from the first half of the first session (Sess-1-H1), the second half of the first session (Sess-1-H2), the first half of the second session (Sess-2-H1), and the second half of the second session (Sess-2-H2). The data were preprocessed and analyzed using NIRS_SPM 4.0[49] and MATLAB 2012b (The MathWorks Inc., Massachusetts). To obtain relatively stable and low-noise data, the first 34 s and last 3 s of data of each data segment were discarded. The remaining data were then downsampled to 8.77 Hz (half of the sampling rate), and second-order drifts were removed. A discrete cosine transform-based high-pass filter with a cut-off frequency of 0.0078 Hz and a low-pass filter based on the hemodynamic response function (HRF) were applied to remove slow drifts and high frequency noises.

The general linear model was used to detect the hemodynamic activities of HbO, HbR, and HbT signals from each of the four data segments from each participant. The design matrix consisting of three boxcar regressors (two for emotion categories and one for self-report-rating) was convolved with a Gaussian HRF to obtain the predictors of the time series of brain activation. Beta-weights, scaling the predictors, represented the weight of the task to the time series of the hemodynamic response. At the individual level, analyses were carried out channel-by-channel, and the beta-weights were tested by one-sided t test under the unpleasant–neutral contrast. At the group level, a one-sided t test under the unpleasant–neutral contrast was carried out on beta-weights of all participants channel by channel.

2.6 *Test–Retest Reliability Assessment*

The fNIRS test–retest reliability of emotion-related frontal activation was assessed for long (intersession) and short (intrasession) separation times and for the three types of hemoglobin signals (HbO, HbR, and HbT). The intersession reliability was assessed using two pairs of datasets: Sess-1-H1 and Sess-2-H1, as well as Sess-1-H2 and Sess-2-H2. Similarly, the intrasession reliability was assessed using datasets of Sess-1-H1 and Sess-1-H2, as well as Sess-2-H1 and Sess-2-H2. For a comprehensive

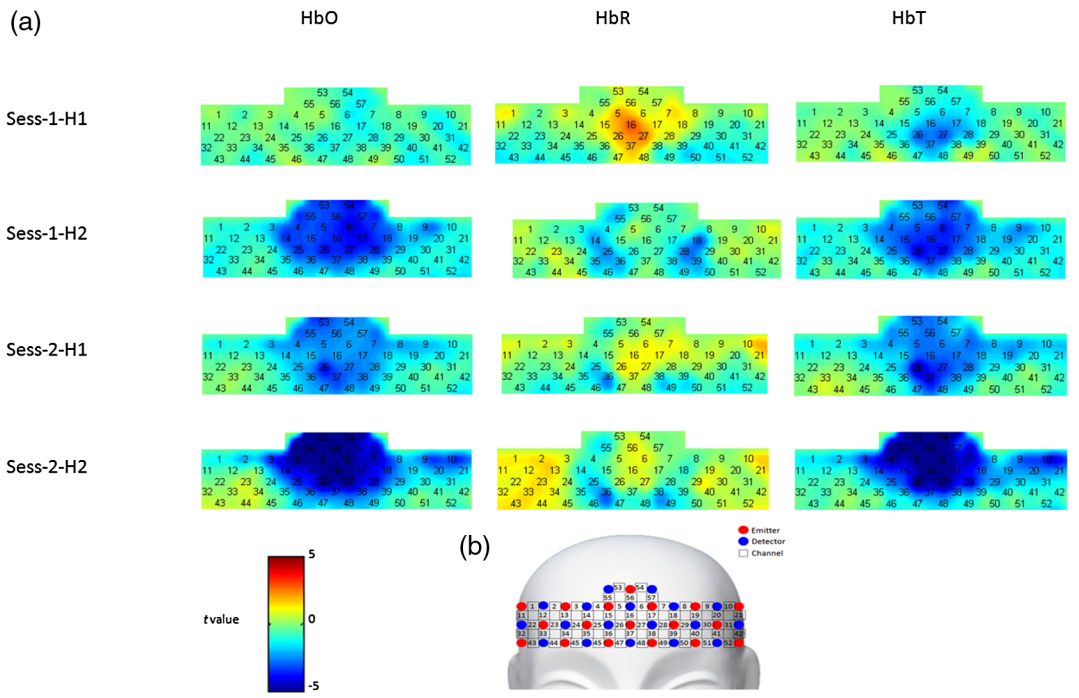

Fig. 2 (a) The activation maps at the group level. (b) Scheme of the fNIRS optodes and channels used in the current study. The center optode in the lowest row was placed at FPz of the International 10–20 system.

evaluation of the reliability, the assessment was conducted on three spatial scales: from a global mapwise scale, to a localized clusterwise scale, then a channelwise scale.

2.6.1 *Mapwise*

For the mapwise assessment, Pearson's correlation coefficient (r) was used to evaluate the similarity of the global activation maps across and within sessions. The mapwise reliability was assessed at both the group and individual levels.

2.6.2 *Clusterwise*

For the clusterwise assessment, two indices, R_{overlap} and an intraclass correlation coefficient ($\text{ICC}_{\text{cluster}}$), were used to assess the reproducibility of frontal activation based on obvious activated clusters under corresponding definitions. R_{overlap}[25,50] measured the degree of spatial overlap of the activated clusters between the test and the retest sessions. The activated clusters consisted of a fixed proportion of channels with the highest t-values.[51] For a comprehensive evaluation, four different proportions were used to define the cluster size: top 10%, top 20%, top 30%, and top 40%. R_{overlap} was calculated by the following equation: $R_{\text{overlap}} = 2 \times C_{\text{overlap}}/(C_1 + C_2)$, where C_1 and C_2 represent the number of activated channels (i.e., the number of channels within the defined cluster) in the test and retest occasions, respectively, and C_{overlap} is the number of the overlapped activated channels across the two sides. The R_{overlap} assessment was performed at both the group and individual levels. $\text{ICC}_{\text{cluster}}$[25,31] was used to quantify the reproducibility of the averaged activation intensity in the activated cluster, which was defined as the activated channels at the group level in the first session (for the intersession assessment) or the first half of each session (for the intrasession assessment). The cluster size was defined as for R_{overlap}, i.e., from the top 10% to 40%. $\text{ICC}_{\text{cluster}}$ was calculated based on a two-way random effect model for consistency measurements.[52] The reliabilities of both single [ICC (C, 1)] and average measures [ICC (C, k)] were evaluated by $\text{ICC}_{\text{cluster}}$, according to the following equations:

$$\text{ICC}(C,1) = \frac{\text{MS}_\text{s} - \text{MS}_\text{e}}{\text{MS}_\text{s} + (k-1)\text{MS}_\text{e}}$$

and

$$\text{ICC}(C,k) = \frac{\text{MS}_\text{s} - \text{MS}_\text{e}}{\text{MS}_\text{s}},$$

where MS_s is the between-subject mean square, MS_e is the error mean square, and k is the number of measurements (in this study $k = 2$).

2.6.3 *Channelwise*

For the channelwise assessment, the $\text{ICC}_{\text{channel}}$ was used to quantify the reproducibility of frontal activation at the single channel level. The calculation method for $\text{ICC}_{\text{channel}}$ was the same as that for the $\text{ICC}_{\text{cluster}}$.

The reliability levels quantified by the indices above were graded according to the criteria proposed by Cicchetti and Sparrow,[53] wherein reliability having a value of >0.75 was considered "excellent," 0.59 to 0.75 as "good," 0.40 to 0.58 as "fair," and <0.40 as "poor." In the fNIRS reliability literature for other psychological functions, a "reliable" or "repeatable" measurement usually has a reliability grade above or close to "fair." For example, in a motor control task, NIRS modulations with Pearson's r from about 0.3 to 0.6 are concluded as repeatable.[27] In a study[28] with a verbal fluency task, the mapwise r values are from 0.55 to 0.83 and the $\text{ICC}_{\text{cluster}}$ values are from

0.26 to 0.78. These levels are regarded as acceptable reliability. The fNIRS-based resting-state functional connectivity in the sensorimotor regions is evaluated as reliable in a test–retest assessment,[31] in which the group-level mapwise r values are from 0.75 to 0.88, the clusterwise $R_{overlap}$ values are from 0.20 to 0.80, and the $ICC_{cluster}$ values are from 0 to 0.69. Thus, in the current study, the reliability value reaching the fair level or above is regarded as an acceptable reliability.

2.7 *Behavioral Data Analysis*

The self-reported PANAS scores and emotional experiences elicited by affective pictures were analyzed using Statistical Package for the Social Sciences (SPSS v.20.0). Two paired-samples t-tests were used to test whether there were significant differences between the test and the retest sessions in positive affect (PA) and negative affect (NA) scores, respectively.

For the first halves of the test and the retest sessions, two 2 (emotion: unpleasant and neutral) × 2 (session: test and retest) repeated measures ANOVAs were carried out to analyze the valence and the arousal scores indicating emotional experiences in the experiment. The same analyses were carried out for the second halves of the test and the retest sessions. Then, for the test and the retest sessions, four 2 (emotion: unpleasant and neutral) × 2 (half: first and second) repeated measures ANOVAs were conducted to compare the differences of emotional experiences between the first and the second halves.

3 Results

3.1 *PANAS Scores and Affective Ratings*

Participants' affective states before the experiment were not significantly different between the test and the retest sessions [PA: $PA_{test} = 32.92 \pm 4.34$, $PA_{retest} = 31.85 \pm 4.39$, $t(25) = 1.12$, $p > 0.05$; NA: $NA_{test} = 17.15 \pm 4.30$, $NA_{retest} = 16.54 \pm 3.60$, $t(25) = 0.85$, $p > 0.05$].

As shown in Table 1, in the emotion × session ANOVAs, analyses with affective ratings found significant main effects of the emotional category (all $ps < 0.001$), indicating that unpleasant pictures induced more unpleasant and more stimulating feelings than neutral pictures. The main effects of session were not significant for either the valence or the arousal dimension (all $ps > 0.05$), indicating that participants' emotional experiences were unchanged in the intersession comparisons. There was no significant emotion × session interaction (all $ps > 0.05$).

In the emotion × half ANOVAs, the statistical results were similar to those of the emotion × session analyses. The emotion main effects for valence and arousal were significant (all $ps < 0.001$). The half effects were not significant (all $ps > 0.05$). There was no significant emotion × half interaction (all $ps > 0.05$).

3.2 *Hemodynamic Responses*

As shown in Fig. 2, at the group level, the HbR concentration had a sign of decrease ($t > 0$) for the unpleasant-neutral contrast at channels 5, 6, 15, 16, 17, 26, 27, and 37, but it was not statistically significant. The localizing analysis indicated that these channels were located above the frontal polar area (Brodmann Area 10 by channels 16, 26, 27, and 37) and the dorsolateral prefrontal area (Brodmann Area 9 by channels 5, 6, 15, and 17). For HbO and HbT, there were nearly no positive t values found at the prefrontal channels. For the three types of hemoglobin signals, it could be seen by visual inspection that the activation maps for the test and retest sessions were similar, and also for the data segments within the same sessions.

3.3 *Test–Retest Reliability*

3.3.1 *Mapwise*

Figure 3 shows the correlations of the group-level activation maps, with Pearson's r values annotated. Each data point in the scatterplots indicates a pair of t values for the test and the retest occasions at the same channel. The mapwise intersession reliabilities at the group level were good to excellent ($r = 0.65$ to 0.88) for all the three signals except that the HbO reliability (Inter-1) for the comparison between Sess-1-H1 and Sess-2-H1 had a poor grade ($r = 0.11$). The mapwise intrasession reliability varied in the test and the retest sessions. In the test session (Intra-1), the HbT signal had an excellent ($r = 0.77$) intrasession reliability, but the reliability for HbO and HbR was poor (up to 0.28). The intrasession reliability in the retest session (Intra-2) was better than that in the test session, with an excellent grade (up to 0.85) for HbO and HbT, and fair grade ($r = 0.46$) for HbR.

The mapwise reliability at the individual level is presented in Table 2. The averaged r values indicated a poor reproducibility

Table 1 The means and the standard deviations (in the parentheses) of self-reported valence and arousal elicited by affective pictures. The F values and significances of the ANOVAs for the intersession comparisons are shown.

	Valence					Arousal				
	Unpleasant	Neutral	E	S	E×S	Unpleasant	Neutral	E	S	E×S
Sess-1-H1	2.07 (0.76)	5.80 (0.79)	343.2 9[a*]	0.26[a]	2.87[a]	6.75 (1.51)	2.77 (1.14)	126.02[a*]	1.48[a]	0.00[a]
Sess-2-H1	2.25 (0.68)	5.70 (0.80)				6.60 (1.54)	2.61 (1.24)			
Sess-1-H2	2.09 (0.88)	5.60 (0.83)	276.74[b*]	2.41[b]	1.10[b]	6.91 (1.61)	2.68 (1.23)	115.87[b*]	2.37[b]	4.02[b]
Sess-2-H2	2.35 (0.82)	5.65 (0.76)				6.46 (1.61)	2.74 (1.12)			

E: main effect of emotion; S: main effect of session; E × S: interaction effect of emotion × session.
[a]F values of the emotion (unpleasant, neutral) × session (Sess-1-H1, Sess-2-H1) ANOVA.
[b]F values of the emotion (unpleasant, neutral) × session (Sess-1-H2, Sess-2-H2) ANOVA.
[*]$p < 0.001$.

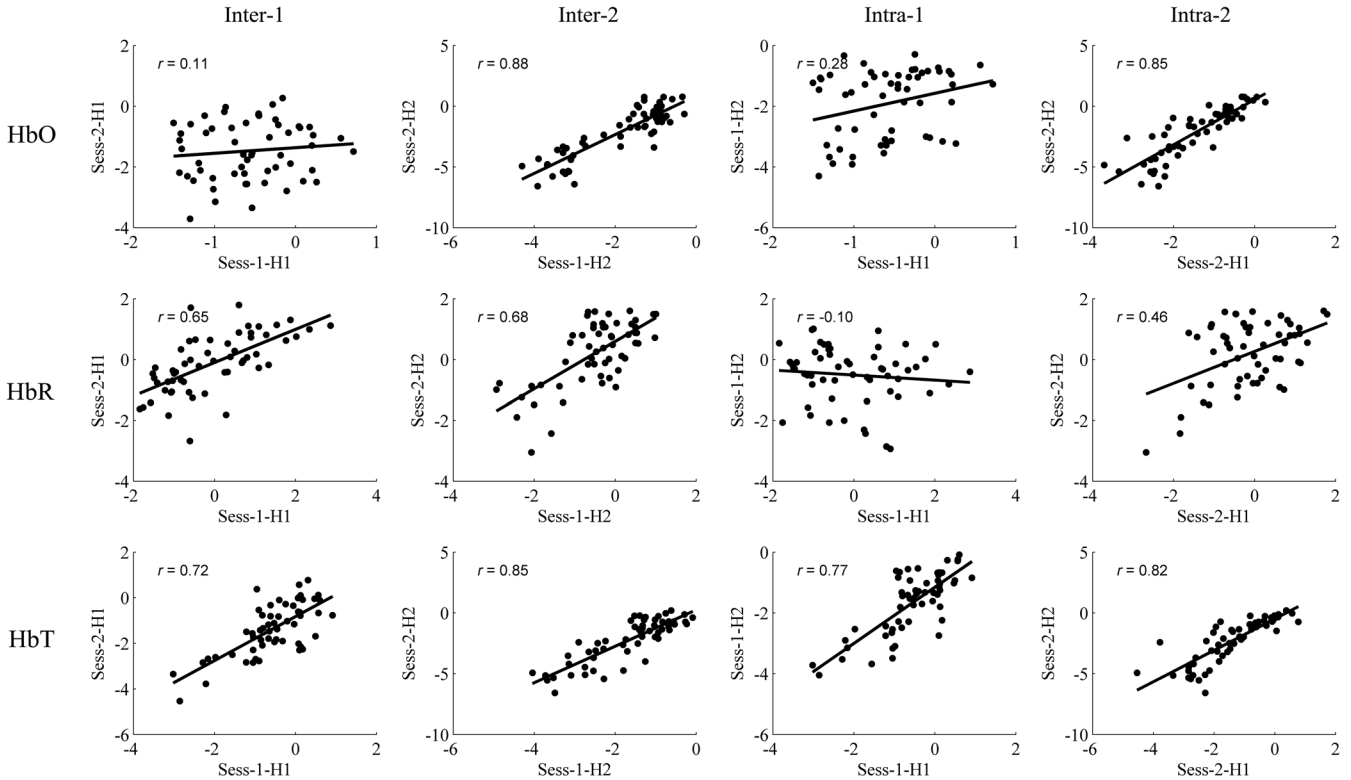

Fig. 3 The scatter plots of map-wise reliability at the group level for HbO, HbR, and HbT signals. Each dot in the scatterplots represents the magnitude of group-level frontal activation (i.e., *t* values) at the same channel. Inter-1 is the intersession reliability for the first halves of two sessions; inter-2 is the intersession reliability for the second halves of two sessions; intra-1 is the intrasession reliability for the first session; intra-2 is the intrasession reliability for the second session.

of activation maps derived from the three types of hemoglobin signals, regardless of the intersession or intrasession assessment. The reliability for HbT seemed slightly higher than that for HbO and HbR. As shown in Table 2, there were great interpersonal differences in the mapwise reliability assessment.

3.3.2 *Clusterwise ($R_{overlap}$)*

The clusterwise reliability based on R_{overlap} values is summarized in Table 3 (group level) and Table 4 (individual level). At the group level, the clusterwise reliability was generally acceptable (up to 0.91, 75% of R_{overlap} values were higher than 0.4), especially for the HbT signal. The reliability had a trend of increase as the cluster size became larger.

Similar to the situation in the mapwise assessment, the individual-level clusterwise reliability was markedly reduced compared with the group level. It varied from "poor" to "fair" as indicated by the averaged R_{overlap} values. Only 50% of averaged R_{overlap} values reached the fair grade. The individual-level reliability also increased as the cluster size was enlarged. Differences across the three hemoglobin signals, or across the intersession and intrasession, were not obvious. Clusterwise reliability with cluster sizes of the top 40% for each participant is shown in Table 2, where large individual differences are present.

3.3.3 *$ICC_{cluster}$ and $ICC_{channel}$*

The clusterwise reliability evaluated by $\text{ICC}_{\text{cluster}}$, with cluster sizes from the top 10% to the top 40%, is listed in Table 5. Table 6 presents the mean and the standard deviation of the channelwise ICCs ($\text{ICC}_{\text{channel}}$, based on *t* values of each channel), which were calculated across all the 57 channels in the optode holder and across the activated channels with different sizes (top 10% to top 40% of channels with highest *t* values). The percentage of channels with ICC $\geq$ 0.4 (fair reliability) among the total interested channels is shown in Table 6.

$\text{ICC}_{\text{cluster}}$ and $\text{ICC}_{\text{channel}}$ by average measures (ICC (C, k)) were higher than by a single measure (ICC (C, 1)). The clusterwise ICCs were slightly higher than the channelwise ICCs. At the two spatial scales, the intrasession reliability was better than the intersession reliability for all the three types of hemoglobin signals. Generally speaking, $\text{ICC}_{\text{cluster}}$ and $\text{ICC}_{\text{channel}}$ derived from HbT signals were higher than those from HbO and HbR signals. Sixty-six percent of HbT $\text{ICC}_{\text{cluster}}$ values were higher than 0.4, while the proportion was only 13% for HbO, and 28% for HbR. Similarly, 63% of HbT $\text{ICC}_{\text{channel}}$ values were higher than 0.4, while the proportion was 10% for HbO, and 25% for HbR.

During the calculation of HbO-derived ICCs, several negative ICC values were yielded, which is theoretically impossible.[54] The reason for negative ICCs remains unclear.[55] Therefore, the negative ICCs were set to zero (i.e., least reliable) in the current study, as suggested in previous studies.[31,56]

4 Discussion

The current study evaluated the fNIRS test–retest reliability of prefrontal activation patterns elicited by unpleasant pictures. The self-reported affective status showed no significant change

Table 2 Mapwise r values and clusterwise $R_{overlap}$ values with cluster sizes of the top 40% of the entire channels for each participant. The averaged Pearson's r values and $R_{overlap}$ values across participants and the standard deviations (in the parentheses) are also shown.

	Inter-1		Inter-2		Intra-1		Intra-2	
Participant	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)
HbO								
1	−0.02	0.39	0.07	0.39	0.10	0.39	0.55	0.61
2	0.33	0.65	0.78	0.74	−0.61	0.17	−0.17	0.26
3	−0.01	0.26	-0.17	0.30	0.66	0.74	0.38	0.65
4	−0.14	0.39	0.03	0.39	−0.03	0.30	0.39	0.52
5	−0.10	0.30	0.12	0.39	−0.08	0.35	0.55	0.57
6	0.27	0.43	0.44	0.57	0.64	0.70	0.52	0.61
7	0.70	0.78	0.36	0.48	0.77	0.74	0.51	0.70
8	0.27	0.52	0.14	0.43	0.15	0.48	−0.11	0.26
9	0.36	0.61	0.88	0.87	0.67	0.65	0.74	0.74
10	0.55	0.48	-0.07	0.30	−0.17	0.35	−0.21	0.35
11	0.68	0.70	0.67	0.74	0.51	0.61	0.85	0.78
12	0.77	0.78	−0.09	0.39	0.77	0.87	−0.15	0.30
13	−0.39	0.22	0.58	0.65	−0.58	0.26	0.48	0.65
14	0.79	0.78	0.17	0.43	0.52	0.65	0.45	0.57
15	−0.17	0.30	0.69	0.70	0.00	0.39	0.77	0.78
16	0.15	0.35	0.65	0.74	−0.01	0.35	0.84	0.91
17	−0.11	0.39	0.72	0.70	0.36	0.57	0.62	0.65
18	−0.26	0.35	-0.02	0.43	0.12	0.39	0.35	0.65
19	−0.21	0.30	0.15	0.43	0.23	0.48	−0.03	0.43
20	0.38	0.57	−0.24	0.22	0.49	0.61	0.04	0.39
21	−0.13	0.35	0.21	0.48	−0.09	0.39	−0.06	0.43
22	−0.62	0.13	0.67	0.83	−0.36	0.22	0.82	0.83
23	−0.49	0.13	0.50	0.70	−0.56	0.09	0.68	0.74
24	0.08	0.61	−0.13	0.22	0.19	0.57	0.56	0.78
25	−0.07	0.43	−0.51	0.26	−0.19	0.30	−0.57	0.26
26	0.12	0.57	0.09	0.48	−0.02	0.48	0.28	0.52
Mean (Std)	0.10 (0.39)	0.45 (0.19)	0.26 (0.37)	0.51 (0.19)	0.13 (0.41)	0.47 (0.19)	0.35 (0.39)	0.57 (0.19)
HbR								
1	−0.32	0.30	−0.47	0.13	0.42	0.61	0.56	0.61
2	0.25	0.48	0.66	0.65	−0.57	0.17	−0.38	0.30
3	−0.14	0.39	−0.23	0.26	0.29	0.52	0.27	0.48

Table 2 (*Continued*).

	Inter-1		Inter-2		Intra-1		Intra-2	
Participant	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)
4	0.01	0.43	0.48	0.65	0.13	0.43	−0.51	0.22
5	0.18	0.48	0.12	0.52	0.41	0.52	0.64	0.74
6	0.29	0.61	−0.04	0.30	0.00	0.43	0.46	0.70
7	−0.22	0.26	0.02	0.52	0.25	0.48	0.83	0.87
8	−0.15	0.39	0.21	0.48	0.43	0.65	−0.28	0.17
9	0.11	0.43	0.20	0.61	0.47	0.61	0.09	0.43
10	0.10	0.57	0.19	0.43	−0.04	0.43	0.48	0.61
11	0.68	0.74	0.44	0.48	0.82	0.61	0.80	0.65
12	0.48	0.57	0.28	0.57	0.38	0.65	0.10	0.43
13	0.11	0.43	0.07	0.43	−0.08	0.30	0.25	0.57
14	0.23	0.52	0.23	0.57	−0.12	0.35	0.13	0.48
15	0.49	0.61	0.38	0.52	0.20	0.52	0.65	0.61
16	0.40	0.57	0.55	0.70	0.46	0.65	0.51	0.57
17	0.06	0.43	0.47	0.61	0.19	0.48	0.53	0.65
18	−0.13	0.35	0.13	0.48	−0.22	0.35	0.67	0.61
19	0.31	0.48	−0.53	0.26	0.12	0.52	0.54	0.43
20	0.05	0.35	0.47	0.61	0.36	0.65	0.26	0.43
21	−0.08	0.39	−0.01	0.30	0.34	0.52	0.30	0.57
22	−0.21	0.22	0.70	0.78	−0.37	0.22	−0.05	0.48
23	0.29	0.48	0.43	0.57	0.66	0.70	0.47	0.57
24	0.62	0.61	−0.24	0.22	0.68	0.65	0.05	0.39
25	−0.47	0.09	0.21	0.57	0.04	0.43	−0.18	0.30
26	−0.12	0.43	−0.20	0.43	0.21	0.43	0.08	0.35
Mean (Std)	0.11 (0.29)	0.45 (0.14)	0.17 (0.32)	0.49 (0.16)	0.21 (0.32)	0.50 (0.14)	0.28 (0.36)	0.51 (0.16)
HbT								
1	0.40	0.65	0.39	0.70	0.33	0.57	0.50	0.57
2	0.18	0.52	0.43	0.57	−0.22	0.35	0.07	0.48
3	−0.20	0.22	−0.04	0.35	0.60	0.70	0.49	0.65
4	−0.37	0.22	0.02	0.35	0.14	0.43	0.47	0.57
5	−0.13	0.30	0.25	0.57	0.04	0.30	0.64	0.78
6	0.50	0.74	0.39	0.52	0.33	0.48	0.60	0.70
7	0.58	0.65	−0.09	0.39	0.83	0.87	0.33	0.52

Table 2 (*Continued*).

	Inter-1		Inter-2		Intra-1		Intra-2	
Participant	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)	r	$R_{overlap}$ (40%)
8	0.39	0.65	0.08	0.43	0.11	0.48	−0.02	0.35
9	0.21	0.57	0.80	0.78	0.72	0.65	0.61	0.74
10	0.36	0.48	−0.19	0.30	−0.06	0.43	0.17	0.43
11	0.73	0.70	0.68	0.78	0.77	0.70	0.92	0.83
12	0.61	0.70	−0.15	0.35	0.67	0.74	−0.23	0.17
13	0.10	0.39	0.23	0.48	−0.13	0.26	0.33	0.52
14	0.70	0.70	0.04	0.43	0.10	0.39	0.06	0.48
15	0.26	0.43	0.61	0.70	0.35	0.57	0.78	0.70
16	−0.05	0.26	0.65	0.70	0.05	0.48	0.85	0.91
17	0.15	0.43	0.76	0.65	0.58	0.61	0.63	0.70
18	0.09	0.39	0.01	0.52	−0.15	0.35	0.61	0.70
19	0.11	0.26	−0.04	0.35	0.23	0.48	−0.07	0.30
20	0.30	0.48	0.23	0.43	0.55	0.57	0.07	0.39
21	−0.13	0.35	0.09	0.52	0.04	0.39	0.14	0.48
22	−0.15	0.22	0.00	0.35	0.28	0.43	0.81	0.87
23	0.09	0.52	0.70	0.78	0.20	0.52	0.82	0.83
24	0.66	0.83	0.60	0.70	0.66	0.65	0.65	0.83
25	−0.02	0.26	−0.59	0.13	−0.16	0.26	−0.49	0.22
26	0.02	0.48	0.47	0.74	0.06	0.48	0.28	0.52
Mean (Std)	0.21 (0.30)	0.48 (0.18)	0.24 (0.35)	0.52 (0.18)	0.27 (0.32)	0.51 (0.15)	0.39 (0.37)	0.59 (0.20)

across the test and retest sessions, and across the first and the second halves. Both the intersession and intrasession reliabilities of emotion-related prefrontal activation were generally acceptable, especially at the group level. The highest mapwise reliability at the group level (0.88) appeared in the intersession comparison between second half of each session for HbO signal. And in the clusterwise assessment, the highest group-level reliability (0.91) was the $R_{overlap}$ (with a top 40% cluster size) of the intrasession comparison within the retest session for HbT signal. The reliability varied with the assessment scale used. Better stabilities were observed on bigger spatial scales. For different hemoglobin signals, the fNIRS reproducibility also showed some differences. The HbT signal had the best stability among the three types of hemoglobin signal tested.

It may be supposed that the intersession reliability would be worse than the intrasession reliability because the stabilities of equipment and participants' mental status are more likely to show changes over relatively longer time intervals. Furthermore, optode displacement errors between the test and retest sessions may also cause a decrease in reliability. Indeed, there was a report[31] that the fNIRS resting-state functional connectivity in the sensorimotor areas was more stable for intrasession assessment relative to intersession assessment. However, in the current study, the intersession versus intrasession contrasts were not fully consistent across the assessment levels, spatial scales, signal types, or assessment indices. Generally, the intersession and intrasession reliabilities in this study did not show obvious differences when compared. Therefore, emotion-related fNIRS patterns might maintain a similar stability over short and long intervals as set in the current study. However, reliability may be influenced by a variety of factors, some of which are hard to measure and control. For example, in the current case, the first and the second half of the same session used different pictures to elicit emotions. Although the self-reports indicated that the two halves of pictures did not have significant differences in the valence and the arousal dimensions, the stimuli differences due to other unknown or uncontrolled aspects could still potentially decrease the intrasession reliability. Thus, the current study did not find a supposed lower reliability in the intersession assessment relative to the intrasession situation. The influence of temporal interval on fNIRS stability and its underlying factors have rarely been addressed in previous studies. More empirical evidence is still needed to understand this issue better.

Table 3 Clusterwise reliability indicated by $R_{overlap}$ values at the group level. The $R_{overlap}$ values with cluster sizes from the top 10% to top 40% are shown.

		Inter-1	Inter-2	Intra-1	Intra-2
HbO	Top 10%	0.00	0.33	0.17	0.83
	Top 20%	0.00	0.50	0.42	0.67
	Top 30%	0.33	0.56	0.50	0.83
	Top 40%	0.48	0.70	0.57	0.87
HbR	Top 10%	0.50	0.50	0.17	0.50
	Top 20%	0.58	0.83	0.25	0.58
	Top 30%	0.67	0.67	0.22	0.39
	Top 40%	0.78	0.65	0.30	0.43
HbT	Top 10%	0.17	0.17	0.50	0.50
	Top 20%	0.50	0.42	0.42	0.75
	Top 30%	0.61	0.72	0.67	0.83
	Top 40%	0.70	0.70	0.70	0.91

Table 4 Clusterwise reliability at the individual level. The averaged $R_{overlap}$ values across participants and the standard deviations (in the parentheses) are shown. The cluster sizes were from the top 10% to top 40%.

		Inter-1	Inter-2	Intra-1	Intra-2
HbO	Top 10%	0.19 (0.20)	0.21 (0.16)	0.23(0.22)	0.18 (0.16)
	Top 20%	0.31 (0.22)	0.34 (0.18)	0.32 (0.19)	0.38 (0.22)
	Top 30%	0.40 (0.22)	0.42 (0.20)	0.40 (0.21)	0.50 (0.20)
	Top 40%	0.45 (0.19)	0.51 (0.19)	0.47 (0.19)	0.57 (0.19)
HbR	Top 10%	0.13 (0.19)	0.25 (0.19)	0.25 (0.22)	0.29 (0.25)
	Top 20%	0.23 (0.18)	0.32 (0.17)	0.31 (0.15)	0.36 (0.21)
	Top 30%	0.36 (0.16)	0.42 (0.18)	0.43 (0.15)	0.45 (0.17)
	Top 40%	0.45 (0.14)	0.49 (0.16)	0.50 (0.14)	0.51 (0.16)
HbT	Top 10%	0.20 (0.23)	0.19 (0.17)	0.23 (0.23)	0.22 (0.17)
	Top 20%	0.33 (0.21)	0.33 (0.17)	0.34 (0.19)	0.41 (0.23)
	Top 30%	0.41 (0.19)	0.44 (0.18)	0.41 (0.17)	0.52 (0.24)
	Top 40%	0.48 (0.18)	0.52 (0.18)	0.51 (0.15)	0.59 (0.20)

The results of the current study showed that the spatial activation patterns were generally more stable at the group level than at the individual level, which is consistent with previous reports[25,28,31,57,58] of fNIRS reliability on other psychological functions. Group analysis, to a certain extent, reduced the influence of optode displacement error and other occasional factors on the test–retest reliability. Therefore, in terms of reliability, group-level analysis is more worthy of attention in future studies.

In line with previous findings,[28,31,58] the spatial scale obviously influenced the reproducibility of the fNIRS results. In the group analyses, the reliabilities were largely located in the "fair" to "excellent" range for the mapwise scale and the clusterwise scales, while for the channelwise scale, the reliabilities were from "poor" to "fair." Therefore, cluster and larger spatial scales can guarantee an acceptable reliability, whereas this is not the case for the channelwise scale. Zhang et al.[31] suggested that the reasons for the low reliability of the channel scale may be the optode placement variability between different sessions, individual differences in skull and brain anatomy, and the low spatial resolution of fNIRS. Analyses using relatively larger spatial scales, such as the map scale and the cluster scale, can help to decrease the optode displacement error and to cancel out individual differences. They suggested that a cluster of fNIRS channels should be treated as a minimal analytical unit when interpreting and comparing the results of fNIRS-based resting-state functional connectivity. The current study also revealed that, for the sake of reliability, the cluster can be used as a minimal spatial unit when studying the fNIRS-based cortical responses to emotional stimuli.

In the current study, the HbT signal was most stable among the three types of hemoglobin signal. At the group level, HbT usually had "fair" to "excellent" reliabilities. It is noteworthy that, although the fNIRS-based neurocognitive studies aimed to measure the functional hemodynamic responses at the outer cerebral cortex, the fNIRS results are inevitably contaminated by hemodynamic fluctuations occurring in the structures (such as the scalp and the pial) above the cortex. These physiological noises will affect the stability of fNIRS results to a certain extent. The tolerance of different types of hemoglobin signal to shallow layer interference is different. Kirilina et al.[59] measured the frontal lobe activation during a continuous performance task and reported that task-evoked scalp vessel artifacts were mainly observed in the HbO concentration changes. Gagnon et al.[60] estimated the cortical contribution to the fNIRS signals in a finger-tapping task. Their results suggested that, relative to HbO and HbR, the HbT signal change was far less sensitive to pial vein influence. Perhaps partly because HbT has a better antinoise ability, we found that it had the best stability in the current study. However, regarding the reliability of different kinds of near-infrared signals, no specific rule has been reported in the previous literature. For example, in a visual task[25] and a verbal fluency task,[28] the researchers did not find constantly higher reliability for one particular fNIRS signal. Although Zhang et al.[31] reported that the rest–retest reliability was generally in the order of HbO > HbR (HbT reliability was between that of HbO and HbR) in their resting-state functional connectivity analyses, another study[30] on the reliability of graph metrics in functional brain networks found the reliability was concordant across HbO, HbR, and HbT. To answer why the reliabilities of different near-infrared signals are different, and what kind of signal is the most reliable, further research is required.

Although the current study found that the fNIRS-based prefrontal responses to emotional stimuli had acceptable stability at the group level and on relatively big spatial scales, it is worth noting that the prefrontal activation magnitudes for the unpleasant-neutral contrast did not reach a statistical significance. In previous studies similar to the current one, some also reported

Table 5 Clusterwise reliability indicated by intraclass correlation coefficient. The $ICC_{cluster}$ values with cluster sizes from the top 10% to top 40% are shown.

		ICC_{single}				$ICC_{average}$			
		Inter-1	Inter-2	Intra-1	Intra-2	Inter-1	Inter-2	Intra-1	Intra-2
HbO	Top 10%	0.00	0.20	0.34	0.18	0.00	0.34	0.51	0.30
	Top 20%	0.00	0.15	0.26	0.24	0.00	0.26	0.41	0.38
	Top 30%	0.00	0.12	0.29	0.23	0.00	0.22	0.45	0.38
	Top 40%	0.00	0.10	0.29	0.23	0.00	0.18	0.45	0.37
HbR	Top 10%	0.21	0.17	0.25	0.19	0.34	0.29	0.40	0.32
	Top 20%	0.08	0.17	0.40	0.31	0.14	0.28	0.57	0.47
	Top 30%	0.15	0.07	0.40	0.30	0.26	0.14	0.57	0.46
	Top 40%	0.12	0.18	0.35	0.31	0.21	0.31	0.51	0.48
HbT	Top 10%	0.23	0.40	0.53	0.51	0.37	0.57	0.69	0.68
	Top 20%	0.23	0.30	0.45	0.47	0.37	0.46	0.62	0.63
	Top 30%	0.25	0.27	0.45	0.45	0.39	0.42	0.62	0.62
	Top 40%	0.26	0.25	0.44	0.43	0.42	0.39	0.61	0.60

that the unpleasant condition did not differ from the neutral condition significantly for the HbO signal.[12,15,61] And some studies reported a decrease of HbO concentration when the unpleasant condition was compared to the resting condition in some participants.[24,62] In these studies and the current one, participants were induced into the desired emotion evidenced by the behavioral data. It is unclear why we, and others, failed to find an expected difference for the unpleasant-neutral or unpleasant-resting contrast in the brain data. However, we offer some considerations in the following text. First, we suggest that the phenomenon of a decrease in HbO may be understood from the neural inhibition hypothesis.[63–65] Although neuroimaging studies have traditionally focused more on task-induced neural activation, there are also studies using fNIRS[66,67] and fMRI[68] that have found a task-related signal decrease at the prefrontal area. The neural inhibition may be related to attentional modulation processes,[66] which work to meet the task required attentional demand by suppressing the activity in some brain areas and increasing the neural activity in some other areas. Thus, the deactivation in the current study may be a neural response for maintaining attention to unpleasant pictures. Interestingly, there is a recent study[69] that investigated negative BOLD responses to intermittent photic stimulation by integrating information from fNIRS and fMRI. The results from both techniques confirmed a negative hemodynamic response, which was related to a decrease in HbO concomitant to a lower increase in HbR response, corresponding to a decrease in HbT. This pattern is very similar to the current findings. Although we cannot make further inferences because of the differences of experimental tasks and observation sites, in the future, the multimodal approach used in this report seems a useful method for a comprehensive exploration of the negative hemodynamic responses. The second possible explanation to the deactivation may be related to the default network[70] that consists of the medial prefrontal cortex, the rostral anterior cingulate, the lateral frontal cortex, and some regions of the parietal and the temporal cortex. The instruction to the participants required them to immerse in the presented scenes but did not provide specific instructions of how to elicit the affective states. Some participants might perform the task by calling on some self-related events or memories. These self-referential processes possibly account for the result of deactivation. Third, it may be possible that the superficial noise mentioned above blurred the prefrontal response that was meant to be observed. From the fNIRS emotion literature, it can be seen that few studies have taken effective measures in data acquisition and processing steps to reduce superficial noise that can bias the results. We think that in future studies, the short channel method,[71] i.e., adding additional short source–detector distance optodes to measure the superficial hemodynamic fluctuations and to regress out the superficial noise in subsequent analyses, may help to get purer cerebral responses.

Several other aspects should be taken into consideration in addition to the issues discussed above. First, we used the fNIRS reliability of prefrontal responses to unpleasant pictures as a representative of the fNIRS reliability in emotion studies. However, emotion research involves very complex and diverse situations. For example, there may be differences in the neural basis of pleasant and unpleasant emotions.[72] There may also be differences between emotion induction and emotion regulation.[73,74] FNIRS stability under specific research questions needs more experimental support. In addition, some fNIRS-based emotion studies have selected some of the sensory areas, such as the visual cortex[75] and the auditory cortex,[76] as regions of interest. The reliability and validity in these cases also need to be tested. Second, in the current study, the reliability of the intersession was similar to that of the intrasession, but a previous study[31] showed that the reliability of the intersession

Table 6 Channelwise reliability indicated by intraclass correlation coefficient. The $ICC_{channel}$ values were calculated across all the 57 channels (whole) and across the activated channels with different sizes (top 10% to top 40%). The means (the values outside the parentheses) and the standard deviations (the first value in parentheses) across the interested channels are shown. The percentages of channels with ICC $\geq$ 0.4 among the total interested channels (the second value in parentheses) are also reported.

		Inter-1	Inter-2	Intra-1	Intra-2
ICC_{single}					
HbO	Whole	0.07 (0.11, 2%)	0.22 (0.15, 11%)	0.22 (0.14, 12%)	0.37 (0.18, 44%)
	Top 10%	0.10 (0.12, 0%)	0.22 (0.17, 17%)	0.37 (0.20, 50%)	0.20 (0.06, 8%)
	Top 20%	0.06 (0.09, 0%)	0.19 (0.14, 8%)	0.30 (0.16, 25%)	0.22 (0.09, 12%)
	Top 30%	0.09 (0.11, 0%)	0.17 (0.13, 6%)	0.27 (0.17, 22%)	0.23 (0.10, 13%)
	Top 40%	0.08 (0.12, 0%)	0.16 (0.14, 8%)	0.25 (0.16, 17%)	0.24 (0.13, 9%)
HbR	Whole	0.13 (0.14, 3%)	0.20 (0.14, 10%)	0.30 (0.17, 25%)	0.29 (0.15, 25%)
	Top 10%	0.19 (0.23, 17%)	0.26 (0.14, 17%)	0.27 (0.22, 17%)	0.29 (0.20, 50%)
	Top 20%	0.14 (0.19, 8%)	0.21 (0.13, 8%)	0.33 (0.19, 25%)	0.33 (0.19, 50%)
	Top 30%	0.17 (0.18, 11%)	0.21 (0.11, 11%)	0.33 (0.17, 28%)	0.32 (0.17, 44%)
	Top 40%	0.16 (0.16, 9%)	0.21 (0.12, 9%)	0.31 (0.19, 26%)	0.33 (0.16, 43%)
HbT	Whole	0.19 (0.15, 8%)	0.29 (0.16, 24%)	0.32 (0.15, 25%)	0.46 (0.15, 68%)
	Top 10%	0.28 (0.17, 17%)	0.42 (0.09, 50%)	0.50 (0.15, 67%)	0.49 (0.09, 83%)
	Top 20%	0.30 (0.14, 17%)	0.32 (0.16, 33%)	0.44 (0.18, 50%)	0.46 (0.09, 67%)
	Top 30%	0.24 (0.16, 11%)	0.29 (0.17, 28%)	0.40 (0.17, 39%)	0.43 (0.10, 56%)
	Top 40%	0.25 (0.18, 17%)	0.27 (0.16, 22%)	0.38 (0.17, 39%)	0.41 (0.12, 52%)
$ICC_{average}$					
HbO	Whole	0.11 (0.17, 10%)	0.34 (0.20, 43%)	0.34(0.19, 39%)	0.52(0.20, 67%)
	Top 10%	0.17 (0.19, 17%)	0.34 (0.22, 33%)	0.51(0.24, 83%)	0.33(0.08, 17%)
	Top 20%	0.10 (0.15, 8%)	0.30 (0.18, 22%)	0.44(0.20, 67%)	0.35(0.12, 33%)
	Top 30%	0.14 (0.18, 17%)	0.27 (0.18, 22%)	0.40(0.22, 61%)	0.36(0.14, 39%)
	Top 40%	0.13 (0.18, 17%)	0.25 (0.20, 22%)	0.37(0.21, 52%)	0.37(0.16, 39%)
HbR	Whole	0.20 (0.21, 19%)	0.31 (0.19, 29%)	0.44(0.20, 61%)	0.43(0.19, 56%)
	Top 10%	0.27 (0.30, 33%)	0.40 (0.16, 33%)	0.39(0.25, 50%)	0.42(0.26, 50%)
	Top 20%	0.20 (0.26, 25%)	0.33 (0.16, 25%)	0.47(0.21, 75%)	0.47(0.23, 58%)
	Top 30%	0.25 (0.25, 28%)	0.34 (0.14, 28%)	0.48(0.19, 72%)	0.46(0.22, 61%)
	Top 40%	0.24 (0.23, 22%)	0.33 (0.16, 30%)	0.44(0.22, 65%)	0.47(0.20, 65%)
HbT	Whole	0.30 (0.20, 30%)	0.43 (0.20, 52%)	0.47(0.18, 68%)	0.62(0.15, 95%)
	Top 10%	0.41 (0.23, 67%)	0.59 (0.08, 100%)	0.66(0.13, 100%)	0.65(0.09, 100%)
	Top 20%	0.44 (0.18, 67%)	0.46 (0.20, 67%)	0.59(0.18, 83%)	0.63(0.08, 100%)
	Top 30%	0.37 (0.24, 57%)	0.42 (0.21, 56%)	0.55(0.17, 78%)	0.59(0.10, 100%)
	Top 40%	0.36 (0.24, 57%)	0.40 (0.21, 52%)	0.53(0.18, 74%)	0.57(0.13, 96%)

was higher. Because these two studies used different experimental tasks and analytical methods, it may not be appropriate to compare them. At present, there is little research regarding the relationship between time interval and reliability. In future studies, to set multiple time spans (e.g., a few hours, weeks, or months) in one study will help to explore this issue. Meanwhile, we should fully consider the consistency of the stimulation materials and the optode placement between the test and the retest sessions, and the psychological and the neural habituation along with the time going, which will help us understand further the role of time interval in reliability. Third, this study showed that the HbT signal, relative to HbO and HbR, was more stable, but we still cannot suggest that future emotion studies with fNIRS should rely more on the HbT signal. Given that existing research has not reported consistent results on this issue, it is unclear whether the stability of the three kinds of signal follows a certain rule. The stability may be influenced by various factors, such as experimental tasks, methods of data processing and analysis, NIRS systems or wavelengths used. To draw a more instructive conclusion, we must control the factors that may influence the results, and collect more evidence from similar studies. In addition, it was reported that the spatial specificity of the three types of fNIRS signal might be different.[77] HbR concentration changes were more spatially localized and the spatial specificity of HbT was relatively poor. From the existing findings, to simultaneously analyze all kinds of signals seems to be a more useful approach. Fourth, the validity of fNIRS in emotion studies is still unclear. At present, the fNIRS technique can only directly measure the superficial layer of the cortex; thus, many fNIRS emotion studies have used the prefrontal (especially lateral) cortex as their region of interest, which is a brain area that has a close relationship with emotional activity.[14,78–81] However, according to Pessoa's review,[82] the prefrontal cortex (excluding the orbital and ventromedial areas and the anterior cingulate cortex) is not regarded as one of the core emotional regions when compared with the amygdala and hypothalamus and other brain regions. This may also explain why previous fMRI reliability studies on emotion tend to choose the amygdala as the observation area, rather than the prefrontal lobe. As for the question whether the prefrontal lobe is a suitable area for the observation of emotional activity, or more precisely, what kind of emotional activity can be measured at the prefrontal lobe, relevant theoretical arguments and empirical work will have to continue for some time. In terms of the evaluation of the validity of fNIRS-based prefrontal responses to emotional stimuli, the simultaneous recording of fNIRS and fMRI may help to elucidate this issue.

In conclusion, under the current experimental design and data processing methods, the test–retest reliability of fNIRS-based prefrontal responses to affective pictures was acceptable at the group level for the mapwise and the clusterwise scales, which suggests that the fNIRS technique may be a reliable tool for emotion studies. Meanwhile, caution should be exercised when using the channelwise and individual-level fNIRS results because they may not be sufficiently stable according to the current study.

Disclosures

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This work was supported by the National Basic Research Program of China (Grant No. 2011CB711000) and the National Natural Science Foundation of China (Grant No. 61273287).

References

1. A. Kleinschmidt et al., "Simultaneous recording of cerebral blood oxygenation changes during human brain activation by magnetic resonance imaging and near-infrared spectroscopy," *J. Cereb. Blood Flow Metab.* **16**, 817–826 (1996).
2. S. E. Fox et al., "Neural processing of facial identity and emotion in infants at high-risk for autism spectrum disorders," *Front. Hum. Neurosci.* **7**, 89 (2013).
3. E. Nakato et al., "Distinct differences in the pattern of hemodynamic response to happy and angry facial expressions in infants—a near-infrared spectroscopic study," *Neuroimage* **54**, 1600–1606 (2011).
4. Y. Hoshi and S. J. Chen, "Regional cerebral blood flow changes associated with emotions in children," *Pediatr. Neurol.* **27**, 275–281 (2002).
5. S. B. Perlman et al., "fNIRS evidence of prefrontal regulation of frustration in early childhood," *Neuroimage* **85**, 326–334 (2014).
6. A. Roos et al., "Altered prefrontal cortical function during processing of fear-relevant stimuli in pregnancy," *Behav. Brain Res.* **222**, 200–205 (2011).
7. M. Balconi and C. Cobelli, "rTMS on left prefrontal cortex contributes to memories for positive emotional cues: a comparison between pictures and words," *Neuroscience* **287**, 93–103 (2015).
8. T. Matsubara et al., "Prefrontal activation in response to emotional words in patients with bipolar disorder and major depressive disorder," *Neuroimage* **85**, 489–497 (2014).
9. A. C. Ruocco et al., "Abnormal prefrontal cortical response during affective processing in borderline personality disorder," *Psychiatry Res. Neuroimaging* **182**, 117–122 (2010).
10. L. Gygax et al., "Prefrontal cortex activity, sympatho-vagal reaction and behaviour distinguish between situations of feed reward and frustration in dwarf goats," *Behav. Brain Res.* **239**, 104–114 (2013).
11. T. Muehlemann et al., "In vivo functional near-infrared spectroscopy measures mood-modulated cerebral responses to a positive emotional stimulus in sheep," *Neuroimage* **54**, 1625–1633 (2011).
12. M. J. Herrmann, A. C. Ehlis, and A. J. Fallgatter, "Prefrontal activation through task requirements of emotional induction measured with NIRS," *Biol. Psychol.* **64**, 255–263 (2003).
13. A. Kochel, F. Schongassner, and A. Schienle, "Cortical activation during auditory elicitation of fear and disgust: a near-infrared spectroscopy (NIRS) study," *Neurosci. Lett.* **549**, 197–200 (2013).
14. K. Marumo et al., "Gender difference in right lateral prefrontal hemodynamic response while viewing fearful faces: a multi-channel near-infrared spectroscopy study," *Neurosci. Res.* **63**, 89–94 (2009).
15. H. Y. Yang et al., "Gender difference in hemodynamic responses of prefrontal area to emotional stress by near-infrared spectroscopy," *Behav. Brain Res.* **178**, 172–176 (2007).
16. R. Aoki et al., "Correlation between prefrontal cortex activity during working memory tasks and natural mood independent of personality effects: an optical topography study," *Psychiatry Res. Neuroimaging* **212**, 79–87 (2013).
17. J. Kopf et al., "The effect of emotional content on brain activation and the late positive potential in a word n-back task," *PLoS One* **8**, e75598 (2013).
18. H. Sato et al., "Replication of the correlation between natural mood states and working memory-related prefrontal activity measured by near-infrared spectroscopy in a German sample," *Front. Hum. Neurosci.* **8**, 37 (2014).
19. L. H. Ernst et al., "Prefrontal activation patterns of automatic and regulated approach-avoidance reactions—A functional near-infrared spectroscopy (fNIRS) study," *Cortex* **49**, 131–142 (2013).
20. T. T. Brink et al., "The role of orbitofrontal cortex in processing empathy stories in 4-to 8-year-old children," *Front. Psychol.* **2**, 80 (2011).
21. S. Schneider et al., "Show me how you walk and I tell you how you feel—a functional near-infrared spectroscopy study on emotion perception based on human gait," *Neuroimage* **85**, 380–390 (2014).

22. S. M. H. Hosseini et al., "Decoding what one likes or dislikes from single-trial fNIRS measurements," *Neuroreport* **22**, 269–273 (2011).
23. S. Moghimi et al., "Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy," *J. Neural Eng.* **9**, 026022 (2012).
24. K. Tai and T. Chau, "Single-trial classification of NIRS signals during emotional induction tasks: towards a corporeal machine interface," *J. NeuroEng. Rehabil.* **6**, 39 (2009).
25. M. M. Plichta et al., "Event-related functional near-infrared spectroscopy (fNIRS): are the measurements reliable?," *Neuroimage* **31**, 116–124 (2006).
26. H. Sato et al., "Within-subject reproducibility of near-infrared spectroscopy signals in sensorimotor activation after 6 months," *J. Biomed. Opt.* **11**, 014021 (2006).
27. G. Strangman et al., "Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: test-retest reliability and review of the literature," *Arch. Phys. Med. Rehabil.* **87**, 12–19 (2006).
28. M. Schecklmann et al., "Functional near-infrared spectroscopy: a long-term reliable tool for measuring brain activity during verbal fluency," *Neuroimage* **43**, 147–155 (2008).
29. A. Watanabe et al., "Cerebrovascular response to cognitive tasks and hyperventilation measured by multi-channel near-infrared spectroscopy," *J. Neuropsychiatry Clin. Neurosci.* **15**, 442–449 (2003).
30. H. J. Niu et al., "Test-retest reliability of graph metrics in functional brain networks: a resting-state fNIRS study," *PLoS One* **8**, e72425 (2013).
31. H. Zhang et al., "Test-retest assessment of independent component analysis-derived resting-state functional connectivity based on functional near-infrared spectroscopy," *Neuroimage* **55**, 607–615 (2011).
32. I. Lipp et al., "Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses," *Neuroimage* **86**, 335–342 (2014).
33. M. M. Plichta et al., "Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery," *Neuroimage* **60**, 1746–1758 (2012).
34. B. G. van den Bulk et al., "How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements," *Dev. Cognit. Neurosci.* **4**, 65–76 (2013).
35. M. M. Plichta et al., "Amygdala habituation: a reliable fMRI phenotype," *Neuroimage* **103**, 383–390 (2014).
36. H. Cao et al., "Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state," *Neuroimage* **84**, 888–900 (2014).
37. H. Doi, S. Nishitani, and K. Shinohara, "NIRS as a tool for assaying emotional function in the prefrontal cortex," *Front. Hum. Neurosci.* **7**, 770 (2013).
38. U. Kreplin and S. H. Fairclough, "Activation of the rostromedial prefrontal cortex during the experience of positive emotion in the context of esthetic experience. An fNIRS study," *Front. Hum. Neurosci.* **7**, 879 (2013).
39. S. Oonishi et al., "Influence of subjective happiness on the prefrontal brain activity: an fNIRS study," in *Oxygen Transport to Tissue XXXVI*, Advances in Experimental Medicine and Biology, H. M. Swartz et al., Ed., Vol. **812**, pp. 287–293, Springer Science+Business Media, LLC, New York (2014).
40. R. C. Bendall, P. Eachus, and C. Thompson, "A brief review of research using near-infrared spectroscopy to measure activation of the prefrontal cortex during emotional processing: the importance of experimental design," *Front. Hum. Neurosci.* **10**, 529 (2016).
41. P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): instruction manual and affective ratings," Technical Report A-6, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida (2005).
42. C. L. Sauder et al., "Test-retest reliability of amygdala response to emotional faces," *Psychophysiology* **50**, 1147–1156 (2013).
43. S. B. Manuck et al., "Temporal stability of individual differences in amygdala reactivity," *Am. J. Psychiatry* **164**, 1613–1614 (2007).
44. R. C. Oldfield, "The assessment and analysis of handedness: the Edinburgh inventory," *Neuropsychologia* **9**, 97–113 (1971).
45. L. Huang, T. Z. Yang, and Z. M. Ji, "Applicability of the positive and negative affect scale in Chinese," *Chin. Ment. Health J.* **17**(1), 54–56 (2003).
46. D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect—the Panas scales," *J. Pers. Soc. Psychol.* **54**, 1063–1070 (1988).
47. M. M. Bradley and P. J. Lang, "Measuring emotion—the self-assessment mannequin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry* **25**, 49–59 (1994).
48. D. Tsuzukic and I. Dan, "Spatial registration for functional near-infrared spectroscopy: from channel position on the scalp to cortical location in individual and group analyses," *Neuroimage* **85**(Pt 1), 92–103 (2014).
49. J. C. Ye et al., "NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy," *Neuroimage* **44**, 428–447 (2009).
50. S. A. R. B. Rombouts et al., "Test-retest analysis with functional MR of the activated area in the human visual cortex," *Am. J. Neuroradiol.* **18**, 1317–1322 (1997).
51. C. Tegeler et al., "Reproducibility of BOLD-based functional MRI obtained at 4 T," *Hum. Brain Mapp.* **7**, 267–283 (1999).
52. K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychol. Methods* **1**, 30–46 (1996).
53. D. V. Cicchetti and S. A. Sparrow, "Developing criteria for establishing interrater reliability of specific items—applications to assessment of adaptive-behavior," *Am. J. Ment. Defic.* **86**, 127–137 (1981).
54. V. Rousson, T. Gasser, and B. Seifert, "Assessing intrarater, interrater and test-retest reliability of continuous measurements," *Stat. Med.* **21**, 3431–3446 (2002).
55. R. Muller and P. Buttner, "A critical discussion of intraclass correlation-coefficients," *Stat. Med.* **13**, 2465–2476 (1994).
56. J. Kong et al., "Test-retest study of fMRI signal change evoked by electroacupuncture stimulation," *Neuroimage* **34**, 1171–1181 (2007).
57. M. M. Plichta et al., "Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: reproducibility of activation?," *Hum. Brain Mapp.* **28**, 733–741 (2007).
58. H. Zhang et al., "Is resting-state functional connectivity revealed by functional near-infrared spectroscopy test-retest reliable?," *J. Biomed. Opt.* **16**, 067008 (2011).
59. E. Kirilina et al., "The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy," *Neuroimage* **61**, 70–81 (2012).
60. L. Gagnon et al., "Quantification of the cortical contribution to the NIRS signal over the motor cortex using concurrent NIRS-fMRI measurements," *Neuroimage* **59**, 3933–3940 (2012).
61. S. Ozawa, G. Matsuda, and K. Hiraki, "Negative emotion modulates prefrontal cortex activity during a working memory task: a NIRS study," *Front. Hum. Neurosci.* **8**, 46 (2014).
62. Y. Hoshi et al., "Recognition of human emotions from cerebral blood flow changes in the frontal region: a study with event-related near-infrared spectroscopy," *J. Neuroimaging* **21**, e94–e101 (2011).
63. K. J. Mullinger et al., "Evidence that the negative BOLD response is neuronal in origin: a simultaneous EEG-BOLD-CBF study in humans," *Neuroimage* **94**, 263–274 (2014).
64. A. T. Smith, A. L. Williams, and K. D. Singh, "Negative BOLD in the visual cortex: evidence against blood stealing," *Hum. Brain Mapp.* **21**, 213–220 (2004).
65. A. R. Wade, "The negative BOLD signal unmasked," *Neuron* **36**, 993–995 (2002).
66. G. Matsuda and K. Hiraki, "Sustained decrease in oxygenated hemoglobin during video games in the dorsal prefrontal cortex: a NIRS study of children," *Neuroimage* **29**, 706–711 (2006).
67. S. Shimada et al., "Decrease in prefrontal hemoglobin oxygenation during reaching tasks with delayed visual feedback: a near-infrared spectroscopy study," *Cognit. Brain Res.* **20**, 480–490 (2004).
68. V. D. Calhoun et al., "Different activation dynamics in multiple neural systems during simulated driving," *Hum. Brain Mapp.* **16**, 158–167 (2002).
69. E. Maggioni et al., "Investigation of negative BOLD responses in human brain through NIRS technique. A visual stimulation study," *Neuroimage* **108**, 410–422 (2015).
70. J. R. Andrews-Hanna, J. Smallwood, and R. N. Spreng, "The default network and self-generated thought: component processes, dynamic control, and clinical relevance," *Ann. N. Y. Acad. Sci.* **1316**, 29–52 (2014).
71. I. Tachtsidis and F. Scholkmann, "False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward," *Neurophoton* **3**, 031405 (2016).

72. M. Viinikainen et al., "Nonlinear relationship between emotional valence and brain activity: evidence of separate negative and positive valence dimensions," *Hum. Brain Mapp.* **31**, 1030–1040 (2010).
73. K. N. Ochsner and J. J. Gross, "The neural bases of emotion and emotion regulation: a valuation perspective," in *Handbook of Emotion Regulation*, J. J. Gross, Ed., pp. 23–42, The Guilford Press, New York (2014).
74. E. Glotzbach et al., "Prefrontal brain activation during emotional processing: a functional near infrared spectroscopy study (fNIRS)," *Open Neuroimaging J.* **5**, 33–39 (2011).
75. M. J. Herrmann et al., "Enhancement of activity of the primary visual cortex during processing of emotional stimuli as measured with event-related functional near-infrared spectroscopy and event-related potentials," *Hum. Brain Mapp.* **29**, 28–35 (2008).
76. M. M. Plichta et al., "Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fNIRS) study," *Neuroimage* **55**, 1200–1207 (2011).
77. M. Suh et al., "Blood volume and hemoglobin oxygenation response following electrical stimulation of human cortex," *Neuroimage* **31**, 66–75 (2006).
78. B. M. Fitzgibbon et al., "Low-frequency brain stimulation to the left dorsolateral prefrontal cortex increases the negative impact of social exclusion among those high in personal distress," *Soc. Neurosci.* **11**, 1–5 (2016).
79. T. Himichi, H. Fujita, and M. Nomura, "Negative emotions impact lateral prefrontal cortex activation during theory of mind: an fNIRS study," *Soc. Neurosci.* **10**, 605–615 (2015).
80. J. C. Motzkin et al., "Ventromedial prefrontal cortex damage alters resting blood flow to the bed nucleus of stria terminalis," *Cortex* **64**, 281–288 (2015).
81. W. M. Pauli, T. E. Hazy, and R. C. O'Reilly, "Expectancy, ambiguity, and behavioral flexibility: separable and complementary roles of the orbital frontal cortex and amygdala in processing reward expectancies," *J. Cognit. Neurosci.* **24**, 351–366 (2012).
82. L. Pessoa, "On the relationship between emotion and cognition," *Nat. Rev. Neurosci.* **9**, 148–158 (2008).

Yuxia Huang is interested in affective and social neuroscience research.

Mengchai Mao is interested in affective and social neuroscience research.

Zong Zhang's research field is the brain imaging technique and methods of data processing and analysis.

Hui Zhou is interested in affective and social neuroscience research.

Yang Zhao's research field is the brain imaging technique and methods of data processing and analysis.

Lian Duan's research field is the brain imaging technique and methods of data processing and analysis.

Ute Kreplin Kreplin is interested in affective and social neuroscience research.

Xiang Xiao's research field is the brain imaging technique and methods of data processing and analysis.

Chaozhe Zhu's research field is the brain imaging technique and methods of data processing and analysis.