Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?

M.M. Plichta, a,* M.J. Herrmann, b C.G. Baehne, a A.-C. Ehlis, a M.M. Richter, a P. Pauli, c and A.J. Fallgatter a

a Department of Psychiatry and Psychotherapy, Laboratory for Psychophysiology and Functional Imaging, University Hospital Wuerzburg, Fuechsleinstraße 15, 97080 Wuerzburg, Germany
b Genomic Imaging, University Hospital of Psychiatry and Psychotherapy Wuerzburg, Germany
c Institute of Psychology, Biological Psychology, Clinical Psychology, and Psychotherapy, University of Wuerzburg, Germany

Received 3 May 2005; revised 6 December 2005; accepted 7 December 2005

The purpose of the present study was to investigate the retest reliability of event-related functional near-infrared spectroscopy (fNIRS). Therefore, isolated functional activation was evoked in the occipital cortex by a periodic checkerboard stimulation. During a 52-channel fNIRS recording, 12 subjects underwent 60 trials of visual stimulation in two sessions. The retest interval was set to 3 weeks. Linear correlations of the contrast t values supplemented by scatter plots, channel-wise intraclass correlation coefficients (ICC) as well as reproducibility indices for the quantity of activated channels (RQUANTITY) and the location (ROVERLAP) of the detected activation were calculated. The results at the group level showed good reliability in terms of the single measure ICCs (up to 0.84) and excellent reproducibility quantified by RQUANTITY and ROVERLAP (up to 96% of the quantity and the location were reproducible), whereas the results at the single subjects’ level were mediocre. Furthermore, the reliability assessed by single measurement ICCs improved if regarded at a cluster level.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Functional near-infrared spectroscopy; NIRS; Event-related; Reproducibility; Test–retest reliability; Time series analysis; Oxygenated haemoglobin; Deoxygenated haemoglobin; Total haemoglobin

Introduction

Like functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS) is a haemodynamic-based technique for the assessment of functional activity in the human brain (e.g. Hoshi, 2003; Obrig and Villringer, 2003; Villringer and Chance, 1997). Based on the tight coupling of neural activity and oxygen delivery (Logothetis and Wandell, 2004), changes in the concentration of oxygenated (O2Hb) and deoxygenated (HHb) haemoglobin are noninvasively measured by fNIRS and taken as indicators for cortical activation. The typical fNIRS signal observed after neural activation is a decrease of HHb accompanied by an increase of O2Hb comparable in time course to the blood oxygenation level dependent (BOLD) signal of fMRI. Compared to the BOLD signal, NIRS provides comprehensive information about haemodynamics consisting of O2Hb, HHb and changes in total haemoglobin (sum of O2Hb and HHb). However, fNIRS measurements are limited to the cortical surface, whereas fMRI and PET enable to image the entire brain (including cortical and subcortical regions).

Because fNIRS utilises the optical rather than radioactive or magnetic properties of the cerebral blood, no contrast agent applications (PET) or complex technical arrangements (e.g. onsite cyclotron (PET), huge magnet (fMRI)) are necessary. Compared to fMRI or PET, fNIRS is characterised by its straightforward application which resembles in the outward appearance more an electroencephalogram (EEG). Thus, the data collection is comfortable for the subjects because of the less constrictive measurement circumstances (e.g. less movement restrictions, no noise disturbance) which probably lead to more ecologically valid conditions than in other neuroimaging methods (e.g. Suzuki et al., 2004). These inherent advantages accompanied by the rapid developments in technology (e.g. from single-channel to multi-channel systems) and methodology (e.g. event-related study designs; time series analysis) enabled fNIRS to easily enter psychological, psychiatric and basic research on children, adults and elderly subjects (e.g. Ehlis et al., 2005; Fallgatter and Strik, 1997, 1998; Fujiwara et al., 2004; Herrmann et al., 2005a, 2005b; Kusaka et al., 2004; Obrig et al., 2000; Safonova et al., 2004; Schroeter et al., 2004b; Shimada et al., 2004, 2005). Despite the enthusiastic use of fNIRS, it seems unmindful that there is little published work focussing on quality criteria of fNIRS in general or the reliability and reproducibility.
of the measured activation in particular. As in the former days of fMRI research, it seems to be tacitly assumed by researchers and practitioners that fNIRS measurements are inherently reliable. Meanwhile, this misinterpretation is overcome in fMRI by means of testing several paradigms regarding the reproducibility of the acquired activation (e.g. Manoach et al., 2001; Smith et al., 2005; Stark et al., 2004; Tegeler et al., 1999; Wei et al., 2004; for further references, see also below). Some efforts have been made to simultaneously combine fNIRS with other (neuroimaging) techniques, and this strategy (in particular, combining fNIRS and fMRI) could be regarded as an approach to validation (e.g. Horovitz and Gore, 2004; Kennan et al., 2002; Kleinschmidt et al., 1996; Mehagnoul Schipper et al., 2002; Murata et al., 2004; Okamoto et al., 2004a; Strangman et al., 2002; Toronov et al., 2001). However, any discrepancies in the results of the combined techniques can either be regarded as measuring different facets of neural activity (e.g. BOLD signal versus O$_2$Hb detection) or simply as a consequence of using instruments with different reliabilities.

In the field of fNIRS research, a block-design study of Watanabe et al. (2003) exists, wherein five subjects were explored two times with 24-channel fNIRS. The resulting reliability was assessed by intraclass correlation coefficients (ICC) for three different performed tasks, which varied from 0.42 to 0.87 (whereas it remains unclear if single or average measure ICCs were reported). The authors conclude that the obtained reliability is acceptable. However, no general conclusions can be drawn from this study because the sample size was small and the retest interval was not held constant for all subjects. A second study from Menke et al. (2003) is focussing on the reproducibility of NIRS parameters measured by two observers in 25 neonates. However, no task or intervention was practiced. Another study (Yoshitani et al., 2002) compares the measurements of two different NIRS instruments during a CO2 challenge test which can be evaluated as a parallel-test study. The two NIRS instruments differed in their used wavelengths, and therefore it is not clear which factors contributed to the partly pronounced differences. Therefore, the latter studies are of no relevance for the present purpose of fNIRS.

To date, no study exists which explores the reliability of event-related fNIRS, which has been recently accomplished and will definitely be a growing field in the application of fNIRS (e.g. Horovitz and Gore, 2004; Kennan et al., 2002; Schroeter et al., 2004a; Tsujimoto et al., 2004). In the present paper, we seek to determine the extent to which event-related fNIRS-measured cortical activations are reproducible under similar conditions over a time period of 3 weeks. In line with fMRI retest studies (e.g. Miki et al., 2001; Rombouts et al., 1997; Specht et al., 2003), isolated functional activation is evoked in the occipital region by presenting a widely used checkerboard as a periodic sensory stimulus in two sessions. The analyses focus on three different key aspects to address the concept of reliability at several levels: first, we investigate each single subject in its variability of the haemodynamic response. Second, a channel-wise comparison is performed in the second level analyses. Third, we compare the group results in an activation map-wise manner. Our aim is to give an overview of the credibility of data derived from fNIRS in event-related experiments as well as some advices for longitudinal case and group studies. Furthermore, we provide some indications for the ongoing debate whether O$_2$Hb or HHb is the more reliable parameter.

Materials and methods

Subjects and task design

Twelve healthy volunteers (seven female and five male, mean age 29.5 ± 6.2 years) participated twice in the present study. The retest interval was set to 3 weeks (mean interval 20.6 days ± 1.8). All subjects had normal or corrected to normal vision and were right-handed. No subject had a known history of any neurologic or psychiatric disorder. All subjects were informed about the nature of the experiment as well as the operating mode of the fNIRS instrument, before giving their written informed consent. The fNIRS investigation of healthy participants was approved by the Ethics Committee of the University of Wuerzburg and by the research conference of the Department of Psychiatry and Psychotherapy.

Subjects were seated in a comfortable chair facing a 21 in. monitor at a distance of approximately 60 cm in a completely dark room. A brief instruction to remain relaxed and to avoid any major body movement was given. The visual stimulation was realised in an event-related paradigm by presenting a simple checkerboard for 1200 ms reversing in contrast at 6 Hz (according to Ozus et al., 2001) followed by 13.8 s of a black screen presentation. Number of trials was set to $n = 60$.

Functional near-infrared spectroscopy

The fundamentals of fNIRS are described in detail elsewhere (e.g. Hoshi, 2003; Obrig and Villringer, 2003). Measurements were performed on a continuous wave system (ETG-4000, Hitachi Medical Co., Japan) using a 3 × 11 optode probe set (consisting of 16 photo-detectors and 17 light emitters) resulting in a total of 52 channels (see Fig. 1A). Two different wavelengths (695 ± 20 nm and 830 ± 20 nm) are used by the system, and its frequency is modulated for wavelengths and channels to prevent crosstalk. Reflected light (not absorbed) leaving the tissue is received by the photo-detectors and transmitted into a set of lock-in amplifiers which are limited to the particular frequencies of interest. Both wavelengths are used to solve the modified Beer–Lambert equation for highly scattering media that allows estimating changes in HHb and O$_2$Hb based on the measurements. Since continuous wave systems cannot measure the optical path length (Hoshi, 2003) and no specific value for the optical path length is adopted from literature (e.g. Duncan et al., 1996), the scale unit is molar concentration multiplied by the unknown path length (mmol × mm). Thus, an absolute quantification is not achieved (systems that enable to measure or estimate the path length are described, e.g. in Duncan et al., 1996 or Firbank et al., 1998). The interoptode distance was 30 mm, which results in measuring approximately 30 mm beneath the scalp. Sampling rate was set to 10 Hz.

The probe set of 52-channel fNIRS was placed on the scalp with its lowest-row centre optode at the subjects’ inion, extending symmetrically towards positions T7 and T8 (not exactly terminating on these positions because of the fixed interoptode distances) according to the international 10–20 system (for the cranio-cerebral correspondence, see Okamoto et al., 2004b). Thus, the EEG positions O1 and O2 were approximately located in the middle-optode row (in-between channels #25/26 and #27/28, respectively). Since the spatial resolution of fNIRS in the axes parallel to the scalp is poor (no better than the emitter–detector distance if non-overlapping geometric arrangements of optodes are
used, see Boas et al., 2004a, 2004b) and no anatomical image was available to take account of the anatomical variance, we defined a spacious region of interest (ROI) including the abovementioned channels plus a surrounding area of 30 mm covering a total of 22 channels. To evaluate a more focussed view, a second ROI was defined, tightly surrounding the channels of interest (covering a total of 8 channels, see Fig. 1A).

Data analysis

The data were analysed using Matlab (The MathWorks Inc., MA, USA) software. To remove baseline drifts and pulsation due to heartbeat, the raw data were pre-processed by applying a band pass filter with cut-off frequencies of 0.02 Hz and 0.7 Hz. Furthermore, we removed trials with artefacts. Because of the lack of established criteria for removing artefact trials from fNIRS datasets, we only removed prominent and abrupt signal changes (defined as signal change of >10% of the baseline level in a time-window of <2 s).

The pre-processed data were then analysed by the two-stage ordinary least squares (OLS) estimation methodology (e.g. Bullmore et al., 1996) according to the general linear model. A two-parameter gamma haemodynamic response function (HRF) was used as a predictor1 for the \(O_2\)Hb, the HHb and the tot-Hb time series (for the rationale of using the HRF for all three fNIRS parameters, see below). At the single subject level, we included the first and second temporal derivative of HRF in order to modulate the onset as well as the dispersion of the HRF. A delta function indicating the onset of sensory stimulation was convolved with the predictors. Thereafter, the first-stage OLS estimation was performed, and the resulting residuals were inspected for model conformity. On the basis of this inspection, we corrected our analyses for serial correlated errors by fitting a first-order autoregressive process to the error term by the Cochrane–Orcutt procedure (Cochrane and Orcutt, 1949). Finally, the beta weights were re-estimated (second stage) and tested for statistical significance by one-sided \(t\) tests (single subject level). For the group analyses, we used a random-effects model with the beta weights of every single subject as the dataset. The statistical inferences at the second level were done by using simple one-sided \(t\) tests. For both single subject and group level, significant cortical activation is indicated by positive \(t\) values for \([O_2Hb]\) as well as \([tot-Hb]\) and by negative \(t\) values for \([HHb]\). To take account of multiple testing, all statistical inferences were based on an adjusted alpha level of 5%. This was realised by using Bonferroni correction and the calculation of the Dubey/Armitage–Parmar alpha boundary (e.g. Sankoh et al., 1997) which include the spatial correlation among the channels. The first session was taken as a frame of reference, and the statistical inferences were made on the basis of its alpha level.

Tests for reproducibility

For testing the reproducibility of cortical activation, three different aspects were focussed on. First, the single subject level was examined. Linear correlations of the individual \(t\) values (sessions 1 and 2) were calculated to assess the amount of similarity in the measured activation of each subject. Second, a channel-wise examination was performed at the group level. To accomplish this, the intraclass correlation coefficient (ICC) was used for quantifying the reliability of the fitted haemodynamic response over the two measurements. For this purpose, the one-way random-effects model was utilised (Shrout and Fleiss, 1979), which implies that the fNIRS measurements are invariant over time and only the ratio of between- and within-subject effects affects the ICC. If the within-subject variation is low over time, the ICC is

---

1 The used HRF is borrowed from the fMRI data analysis. It can be argued that the fMRI-HRF and the NIRS response functions of \(O_2\)Hb, HHb and tot-Hb are different in shape. However, we took this well-known predictor model as a first estimate of the fNIRS signal and accepted a probable misfit. As a matter of course, the best-fitting model function should be investigated in a separate study (cf. Zhang et al., 2005).
close to one. In order to draw conclusions for the single measures as well as for the accuracy of the derived mean value over both sessions, we report the single measure ICC and the average measure ICC (for an example of use, see Johnstone et al., 2005). For the ICC analyses, we inspected the reproducibility of the two ROIs (see above). Additionally, a data-driven approach was applied by defining the activated channels from session 1 as a further ROI (post-hoc view). In order to investigate the reproducibility of mean signal changes across the two sessions, channel-wise and cluster-wise (i.e. mean beta across all channels constituting the ROIs or the post-hoc-view) paired \( t \) tests were performed on the averaged group results, a first visual comparison of both measurements implies a high similarity of the activation pattern across the sessions. To examine the apparent similarity, the single subjects were inspected next.

For the single subjects and the group level analyses, we also compared the quantity of activated channels and the location of the two acquired activation patterns. Therefore, \( R_{\text{QUANTITY}} \) and \( R_{\text{OVERLAP}} \) (according to Rombouts et al., 1997) are calculated:

\[
R_{\text{QUANTITY}} = 1 - \left| A_1 - A_2 \right| / \left( A_1 + A_2 \right) \quad \text{and} \\
R_{\text{OVERLAP}} = 2 \cdot A_{\text{OVERLAP}} / \left( A_1 + A_2 \right)
\]

where \( A_1 \) and \( A_2 \) represent the quantity of the activated channels (operationalised by counting significantly activated channels) of the first and second session, respectively. \( A_{\text{OVERLAP}} \) is the quantity of identical supra-threshold channels in both sessions. Both indices range from 0 (worst) to 1 (best) or can be expressed as percent values, respectively. Since Rombouts et al. (1998) demonstrated a dependency of the overlap index and the used threshold, two different strategies were used for the overlap index: (a) a critical \( t \) value corresponding to the adjusted alpha level of 5% and (b) a fixed number of channels with the highest \( t \) values were inspected (according to Tegeler et al., 1999). Therefore, we explored a top-20%, top-15%, top-10% and top-5% channel quantity threshold.

For an evaluation of the assessed reliability indices (i.e. the retest correlation coefficient, ICCs, \( R_{\text{OVERLAP}} \) and \( R_{\text{QUANTITY}} \)), values of \( \geq 0.80 \) are considered as highly reliable, \( \geq 0.60 \) as moderately reliable and values of \( <0.60 \) as weakly reliable (an overview of different reliability criteria and cut-off values is given by Charter, 2003).

Results and discussion

Figs. 1B–G show the group results (\( t \) maps) derived from both sessions for \( \text{O}_2\text{Hb}, \text{HHb} \) and \( \text{tot-Hb} \) (note that the shown maps are based on interpolations from singular channels). Obviously, fNIRS enables to detect highly significant activation within the occipital region. The evoked neural activity causes a decrease of HHb (Figs. 1D, E) accompanied by an increase of \( \text{O}_2\text{Hb} \) (Figs. 1B, C). Furthermore, tot-Hb increases significantly (Figs. 1F, G). The localisation of the used probe set can be evaluated as adequate, indicated by having the “hot spots” nearby the centre of the observed area. Furthermore, the spatial resolution of fNIRS is sufficient to clearly discriminate the left and right cortices. Based on the averaged group results, a first visual comparison of both measurements implies a high similarity of the activation pattern across the sessions. To examine the apparent similarity, the single subjects are inspected next.

Single subject analyses

Pearson correlation coefficients of the single subjects’ \( t \) values are shown in Table 1. The empirical correlation coefficients for \( \text{O}_2\text{Hb}, \text{HHb} \) and \( \text{tot-Hb} \) indicate moderate (0.63 for \( \text{O}_2\text{Hb} \)) to low (0.53 for \( \text{HHb} \) and 0.49 for \( \text{tot-Hb} \)) reproducibility.

The indices for \( R_{\text{QUANTITY}} \) are moderate for all three fNIRS parameters (\( \text{O}_2\text{Hb} = 0.71, \text{HHb} = 0.70 \) and \( \text{tot-Hb} = 0.64 \)). Thus, the activated areas measured with fNIRS are associated with a comparable quantity of activated channels across the sessions.

Considering the location \( R_{\text{OVERLAP}} \) of the detected activation, the reproducibility is weak. Only 55% (\( \text{O}_2\text{Hb} \)) and 51% (\( \text{tot-Hb} \)) of the

---

**Table 1**

Reproducibility indices at the single subject level (based on the obtained individual \( t \) values)

<table>
<thead>
<tr>
<th>Subject</th>
<th>( \text{O}_2\text{Hb} )</th>
<th>HHb</th>
<th>Tot-Hb</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r )</td>
<td>( R_{\text{QUANTITY}} )</td>
<td>( R_{\text{OVERLAP}} )</td>
<td>( r )</td>
</tr>
<tr>
<td>1</td>
<td>0.84</td>
<td>0.93</td>
<td>0.79 (1.00)</td>
</tr>
<tr>
<td>2</td>
<td>0.41</td>
<td>0.61</td>
<td>0.48 (0.00)</td>
</tr>
<tr>
<td>3</td>
<td>0.43</td>
<td>0.91</td>
<td>0.72 (0.67)</td>
</tr>
<tr>
<td>4</td>
<td>0.57</td>
<td>0.56</td>
<td>0.48 (0.67)</td>
</tr>
<tr>
<td>5</td>
<td>0.56</td>
<td>0.67</td>
<td>0.55 (1.00)</td>
</tr>
<tr>
<td>6</td>
<td>0.76</td>
<td>0.56</td>
<td>0.48 (1.00)</td>
</tr>
<tr>
<td>7</td>
<td>0.47</td>
<td>0.61</td>
<td>0.26 (0.67)</td>
</tr>
<tr>
<td>8</td>
<td>0.84</td>
<td>0.91</td>
<td>0.73 (1.00)</td>
</tr>
<tr>
<td>9</td>
<td>0.84</td>
<td>1.00</td>
<td>0.62 (1.00)</td>
</tr>
<tr>
<td>10</td>
<td>0.39</td>
<td>0.04</td>
<td>0.04 (0.33)</td>
</tr>
<tr>
<td>11</td>
<td>0.84</td>
<td>0.93</td>
<td>0.78 (1.00)</td>
</tr>
<tr>
<td>12</td>
<td>0.66</td>
<td>0.87</td>
<td>0.78 (1.00)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.63</td>
<td>0.72</td>
<td>0.55 (0.78)</td>
</tr>
</tbody>
</table>

---

**Notes:** First value is based on a fixed alpha criterion (5%); values in parenthesis are based on the best fixed number of channels criterion (\( \text{O}_2\text{Hb} = 3 \) channels (top-5%); \( \text{HHb} = 8 \) channels (top-15%); tot-Hb = 10 channels (top-20%)).
significantly activated areas are identical across both sessions. Considering the HHb parameter, this lack of reproducibility is even more pronounced (36% overlap). Interestingly, similar results are reported in fMRI reproducibility studies, where $R_{\text{SIZE}}$ (which is the analogon of $R_{\text{QUANTITY}}$) outperforms the $R_{\text{OVERLAP}}$ index (e.g. Machielsen et al., 2000; Miki et al., 2001; Rombouts et al., 1997, 1998). Even without repositioning (that is, subjects remain in the MRI scanner), Machielsen et al. (2000) showed by applying a visual encoding task that $R_{\text{SIZE}}$ and $R_{\text{OVERLAP}}$ differed considerably ($R_{\text{SIZE}} = 71.6\%$; $R_{\text{OVERLAP}} = 48.9\%$ for the whole brain). The indices decreased to $R_{\text{SIZE}} = 62.2\%$ and $R_{\text{OVERLAP}} = 36.1\%$ (whole brain) after 13.5 days (median) retest interval. The authors concluded that, beside various technical reasons, there is a considerable variance in activation even within subjects.

However, Tegeler et al. (1999) demonstrated that the $R_{\text{OVERLAP}}$ index is lower if a fixed alpha threshold is used compared to a percentage threshold. As shown in Table 1, calculation of $R_{\text{OVERLAP}}$ based on a fixed number of channels leads to improved reproducibility of the location at the single subject level for O2-Hb and HHb ($R_{\text{OVERLAP}} = 78\%$ and $48\%$, respectively) but leads to no substantial change in tot-Hb ($R_{\text{OVERLAP}} = 51\%$).

The reason why there is remaining variability at the single subject level is unknown. In fMRI studies, motion artefacts, repositioning inaccuracy and physiological or psychological changes are discussed as potential contributing factors (e.g. Suzuki et al., 1997, 1998; Tegeler et al., 1999; Wagner et al., 2005). All these factors could also affect the fNIRS measurements. In particular, the accuracy of repositioning the probe set based on anatomical landmarks could be an influential error source and needs further improvement (Sato et al., 2005). To reduce this error, multiple EEG or anatomical marks could be used. Approaches which use MRI scans plus neuronavigational tools (e.g. Suzuki et al., 2004) are also accomplished (ensuring a constant positioning of the probe set based on an anatomical scan), but the inherent advantage that NIRS can be used “on the bedside” would vanish. However, Singh et al. (2005) provide the alternative approach of using a navigational tool without MRI. Another applied approach is to reposition the probe set on the scalp until cortical activation is centred/detected (e.g. Steinbrink et al., 2005). However, there are also system inherent characteristics that further hamper precise localisation (Yamamoto et al., 2002): the current system uses optodes with a diameter of 2 mm and a non-overlapping geometric arrangement of emitters and detectors. As can be seen in Fig. 1A, the measurement points (channels) are presumed to be in the middle of an emitter–detector pair. A punctiform activation located 1 mm to the left of a detector will produce a signal change in the channel located on the left side of this detector. A slight movement of the probe set of, e.g. 3 mm to the left side leads to a detection of the activation spot by the channel located on the right side of the detector. Consequently, there will be a shift of a whole channel position (approximately 3 cm) caused by a movement in the range of millimetres. In practice, no punctiform activation will be existent, but disadvantageous locations of the borders of an activation area could lead to similar channel shifts. This localisation error could occur with any positioning approach if such a border is located close to a detector (even the localisation variability of neuronavigational tools can be up to 2.8 mm, see Schonfeldt Lecuona et al., 2005).

Second level analyses (channel-wise comparison)

The second level analyses for O2-Hb revealed 14 channels (#5, #15–18, #25–29, #35–36, #38–39) with a significant activation in the first session and 13 channels in the second session (#15–18, #25–29, #35–36, #38–39). Considering HHb, four channels (#7, #15, #17, #18) are significantly activated in the first session and three channels in the second session (#15, #17, #26). Analyses of tot-Hb led to the detection of eleven significantly activated channels (#5, #15–18, #25–29, #35) in the first session and 13 channels (#15–18, #25–29, #35–36, #38–39) in the second session. All the significantly activated channels are solely located in the predefined ROI #1 (see Fig. 1A).

From applying the ICC to the group data, it follows that all channels of the predefined ROIs (except channel #37 and #38, $P < 0.10$) demonstrate significant ICCs ($P < 0.05$). The single-channel ICCs and cluster level ICCs (i.e. mean values derived from the constituting channels) of the predefined ROIs as well as a post-hoc view (i.e. the reproducibility of activated channels obtained from session 1) are shown in Table 2. For both predefined ROIs (plus the post-hoc view), the majority of ICCs indicate moderate ($>0.60$) or high reliability ($>0.80$). Eight of 36 ICC indices show weak reliability ($<0.60$) with a pronounced occurrence in the HHb data (four ICCs <0.60) and the tot-Hb data (three ICCs <0.60). The values of HHb are consistently lower or equal to the O2-Hb parameter (one exception is ROI #1, average measure ICC). The values of tot-Hb are the highest (except the single-channel ICCs for the post-hoc view). The cluster level ICCs for all ROIs are higher compared to single-channel ICCs (one exception is the post-hoc view for HHb). Seventeen of 18 ICCs indicate moderate or high reliability.

For [O$_2$Hb], the channel-wise comparisons of the amplitudes lead to the detection of one channel (#50) associated with a

| Table 2 | Reproducibility of the detected activation expressed with intraclass correlation coefficients (ICC) based on the second level parameter set (beta weights) |
|-----------------|-----------------|-----------------|-----------------|
|                | Single channels | Cluster levela  |                |
|                | ICC$_{\text{SINGLE}}$ | ICC$_{\text{AVERAGE}}$ | ICC$_{\text{SINGLE}}$ | ICC$_{\text{AVERAGE}}$ |
| O$_2$Hb        |                   |                   |                |
| ROI #1         | 0.53 (0.08–0.87) | 0.65 (0.15–0.93) | 0.64 | 0.78 |
| ROI #2         | 0.63 (0.32–0.86) | 0.76 (0.49–0.93) | 0.66 | 0.79 |
| Post-hocb      | 0.65 (0.25–0.86) | 0.78 (0.40–0.93) | 0.72 | 0.84 |
| HHb            |                   |                   |                |
| ROI #1         | 0.53 (0.15–0.80) | 0.67 (0.25–0.89) | 0.67 | 0.81 |
| ROI #2         | 0.53 (0.15–0.80) | 0.66 (0.25–0.89) | 0.63 | 0.78 |
| Post-hocb      | 0.52 (0.41–0.59) | 0.68 (0.58–0.75) | 0.52 | 0.69 |
| Tot-Hb         |                   |                   |                |
| ROI #1         | 0.58 (0.09–0.91) | 0.70 (0.16–0.95) | 0.73 | 0.85 |
| ROI #2         | 0.65 (0.25–0.91) | 0.77 (0.41–0.95) | 0.74 | 0.85 |
| Post-hocb      | 0.43 (0.13–0.77) | 0.58 (0.23–0.87) | 0.84 | 0.91 |

a Based on the average of the channels constituting the ROIs and post-hoc area, respectively.

b Based on channels which reach significant results in session 1.
significant difference \((P < 0.05)\). Furthermore, five channels \((\#9, \#36, \#45–\#47)\) demonstrate trends to significance \((P < 0.10)\). For [HHb], two channels \((\#5, \#16)\) are associated with a significant difference \((P < 0.05)\) and one channel \((\#27)\) shows a trend to significance \((P < 0.10)\). Analyses of [tot-Hb] revealed one channel \((\#46)\) associated with a tendency to significance. Regarding the cluster level (ROIs and post-hoc view), no significant differences are detectable (all \(P\) values \((>0.10)\) for \(O_2Hb\), HHb or tot-Hb.

On the basis of these results, it can be stated firstly that single measurement coefficients improve in their reliability if regarded at a cluster level. Secondly, the above-suggested displacement of the used probe set (or any other error) seems to be a random occurrence in the present study (otherwise, the reliability coefficients at the single subjects’ level and the group level would be poor). Thirdly, the number of channels associated with a significant difference \((n = 1\) for \(O_2Hb\), \(n = 2\) for HHb and \(n = 0\) for tot-Hb) is lower than the expected number of false positives by chance if 52 tests for each fNIRS parameter were performed. Therefore, the amplitudes of the haemodynamic responses to the sensory stimulation seem to be stable over the time period of 3 weeks.

Furthermore, the usage of a standard probe set with a fixed number of optodes should be discussed since head sizes are definitely different. At least two problems can arise: first, the statistical power of any applied second level analysis is lowered due to the fact that the inter-individual overlap of the anatomical structure of interest could be very small (or even inexistent) and no normalisation procedure has yet been established. Second, and as a consequence, also the validity is altered because it is common practice in group analyses to average across and focus on channels above an assumed area of interest (mostly based on cranio-cerebral correlations), even if the individual task-related activity is localised elsewhere in individual cases (an alternative strategy is applied in Steinbrink et al., 2005). However, these consequences did not diminish the results of the present retest study because highly significant activation was detectable (indicating that the statistical power was sufficiently high) and the problem of averaging should be equal across the sessions.

**Global comparison (map-wise view)**

As a last step, a global quantification for the similarity of the whole activation maps is extracted. For that purpose, we used the scatter plot as shown in Figs. 2A–C. Each data point represents the \(t\) values of one channel in both sessions. All significantly activated channels are solely located in the predefined ROI #1 (see Fig. 1A). No activation was detectable outside the occipital region. The scatter plot analysis is complemented by the calculation of Pearson correlation coefficient. As can be seen in Fig. 2A, the deviations of the data points from the least squares linear regression line are small for \(O_2Hb\). Regarding \(O_2Hb\), the determination coefficient of 0.80 \((r = 0.89)\) shows that about 80% of the variance is explained by the retest. Finally, the comparison of the quantity and the location for the whole activation maps results in \(R_{\text{QUANTITY}} = 0.96\) which means that 96% of the detected activation in both panels A and C represents the area of the significantly activated channels in both measurements \((P < 0.05,\) corrected). The square in the lower part of panel B represents the area of significantly activated channels for HHb in both measurements. All these channels \((O_2Hb, HHb\) and tot-Hb) are solely located in the predefined ROI #1. Statistics for the second level analysis are shown in the upper left corner of each figure.

![Fig. 2. Scatter plot of the \(t\) values for \(O_2Hb\) (A), HHb (B) and tot-Hb (C) resulting from second level analysis. The upper square in the right corner of panels A and C represents the area of the significant channels in both measurements \((P < 0.05,\) corrected). The square in the lower part of panel B represents the area of significantly activated channels for HHb in both measurements. All these channels \((O_2Hb, HHb\) and tot-Hb) are solely located in the predefined ROI #1. Statistics for the second level analysis are shown in the upper left corner of each figure.](image-url)
sessions are the same regarding the quantity of channels as well as the location).

The map-wise comparison for HHb (Fig. 2B) reveals that the association indicated by the deviation from the regression line is more loose but still sufficient \((r = 0.82)\). The quantity coefficient indicates high reproducibility \((0.86)\), whereas the overlap of the two acquired hot spots is considerably lower than for \(O_2Hb\). Only 57% of the area is reproducible. If the small amount of activated channels is taken into account (four and three for sessions 1 and 2, respectively), a slight variation of the probe set position already leads to a substantial decrement of the overlap coefficient. As can be seen in Fig. 2C, all three tot-Hb reproducibility indices indicate high reproducibility: Pearson correlation coefficient is \(r = 0.86\) (about 74% of the variance is explained by the retest). Furthermore, 92% of the quantity and 83% of the location are identical.

It can be stated that, firstly, these values emphasise the high reproducibility of fNIRS group activation maps in particular for \(O_2Hb\) and tot-Hb. Secondly, \(O_2Hb\) and tot-Hb group results are of better reproducibility compared to HHb. Regarding HHb, the reproducibility is consistently lower. Particularly, if \(R_{\text{OVERLAP}}\) is considered, an obviously lower reproducibility is achieved (57% compared to 96% in \(O_2Hb\) and 83% in tot-Hb) which could mainly be explained by the smaller amount of significantly activated channels.

**Comparison of the fNIRS parameters**

As a matter of course, the question of which parameter is more reliable is highly depending on the particular wavelengths that are used. The ETG-4000 system utilises \(830 \pm 20\) nm (absorption coefficient of \(O_2Hb > HHb\)) and \(695 \pm 20\) nm (absorption coefficient of \(HHb > O_2Hb\)). In contrast, other systems often use a second wavelength of \(>700\) nm. Yamashita et al. (2001) showed that a combination of wavelengths comparable to ours results in measurements which are more precise than measurements with other wavelength pairs (particularly wavelength pairs both \(>700\) nm).

Depending on the used wavelength pair, our results give some indications for the ongoing debate whether \(O_2Hb\) or HHb is a more reliable parameter to measure cortical activation (e.g. Obrig and Villringer, 2003; Schroeter et al., 2002; Strangman et al., 2002). As a consistent finding, it can be stated that HHb was associated with lower \(r\) values at the single subjects’ level as well as at second level if compared to \(O_2Hb\). Thus, only a few channels (in our data, approximately \(1/3\) of the \(O_2Hb\) channel quantity) exhibited significant activation. This can be regarded either as a higher local specificity or as a lack of power to detect significant decrement (otherwise, the reproducibility of \(O_2Hb\) could be evaluated as a consequence of a large and diffuse activation area, where a displacement of the probe is less crucial). Especially for the reproducibility of the location of activation (\(R_{\text{OVERLAP}}\), the narrowed scope associated with HHb was a handicap, whereas the quantity of activated channels is of fair reliability and comparable to the \(O_2Hb\) data indicating an advantage of HHb since it is locally more focussed. The missing reproducibility of the localisation is indeed problematic in the context of a retest study or in a second level analysis. Regarding the ICCs, it can be stated that at the cluster level (ROI #1 and ROI #2) the coefficients associated with HHb are comparable to \(O_2Hb\). Interestingly, the ICCs of tot-Hb outperform the values of \(O_2Hb\) and HHb (except in the posthoc view). This could be due to a reduced crosstalk effect.

However, the remaining reproducibility indices for tot-Hb are in between those for \(O_2Hb\) and HHb (except \(R_{\text{OVERLAP}}\)).

For the practitioner, it can be stated that, if the main focus in data analysis is set on HHb, one should consider that the advantage of a higher local specificity is accompanied with a decline in the reproducibility of the localisation over time and across subjects. This could lead to a seeming signal change across measurement sessions especially if small probe sets are used. Additionally, the second level analyses could be associated with low statistical power. As a consequence, both \(O_2Hb\) and HHb should always be analysed and reported. If they lead to a reasonable accordance, i.e. local decrease of HHb and increase of \(O_2Hb\), it is regarded as an indicator of cortical activation.

**Conclusions**

To our knowledge, this is the first systematic retest reliability study for event-related fNIRS. We used a fixed retest interval of 3 weeks and a sufficiently large sample size of 12 subjects. Based on the above reported results, the following conclusions can be drawn: (1) in single subjects, the degree of reproducibility is highest for \(O_2Hb\) data. The reproducibility of the location can be increased if a fixed number of channels is used as a threshold. In our study, a top-5% criterion yielded the best result for \(O_2Hb\) (mean reproducibility = 78%). An alternative arrangement of optodes could prevent channel shifts and further increase the reproducibility. (2) Group results, in particular for \(O_2Hb\), can be considered as highly stable over time. Regarding \(O_2Hb\), about 96% of the channel quantity and location were reproducible in the present study if a map-wise view was applied. Considering the channel-wise view (ICCs), it can be stated that the reliability is sufficiently high especially when cluster levels are focussed (ICCs for tot-Hb range from 0.73 to 0.84). As in single subjects, it seems advisable not to interpret isolated significant channels which are not framed by or adjacent to other significant channels (which implies that multi-channel instruments are preferable to single-channel systems for functional measurements). Up to now, our results are limited to the occipital lobe. Further brain areas and different paradigms should be examined. (3) Due to the lack of criteria for removing artefact trials from fNIRS datasets, criteria for excluding trials have to be established. In the present study, we proposed one criterion (signal changes of >10% in a time window of ~2 s). Future fNIRS studies should exactly report their observations of potential artefacts to establish a consensus about removing trials or even subjects.

**Acknowledgments**

The authors would like to thank Hitachi Medical Corporation for the ETG-4000 equipment and skilled technical support and Inge Gröbner, Melanie Greutner and Ireen Schaffrath for their proficient technical assistance.

**References**


