Neurophotonics

Neurophotonics.SPIEDigitalLibrary.org

Test–retest reliability of functional near infrared spectroscopy in infants

Anna Blasi
Sarah Lloyd-Fox
Mark. H. Johnson
Clare Elwell

SPIE.

Test–retest reliability of functional near infrared spectroscopy in infants

Anna Blasi,[a,*,†] Sarah Lloyd-Fox,[a,*,†] Mark. H. Johnson,[a] and Clare Elwell[b]
[a]Birkbeck, University of London, Centre for Brain and Cognitive Development, Malet Street, London WC1E 7HX, United Kingdom
[b]University College London, Department of Medical Physics and Bioengineering, Malet Place Engineering Building, Gower Street, WC1E 6BT London, United Kingdom

Abstract. There has been a rapid rise in the number of publications using functional near infrared spectroscopy (fNIRS) for human developmental research over the past decade. However test–retest reliability of this measure of brain activation in infants remains unknown. To assess this, we utilized data from a longitudinal cohort who participated in an fNIRS study on social perception at two age points. Thirteen infants had valid data from two sessions held 8.5 months apart (4 to 8 months and 12 to 16 months). Inter- and intrasession fNIRS test–retest reliability was assessed at the individual and group levels using the oxyhemoglobin (HbO_2) signal. Infant compliance with the study was similar in both sessions (assessed by the proportion of time infants looked to the stimuli), and there was minimal discrepancy in sensor placement over the targeted area between sessions. At the group level, good spatial overlap of significant responses and signal reliability was seen (spatial overlap was 0.941 and average signal change within an region of interest was $r = 0.896$). At participant level, spatial overlap was acceptable (>0.5 on average across infants) although signal reliability varied between participants. This first study of test–retest reliability of fNIRS in infants shows encouraging results, particularly for group-based analysis. [DOI: 10.1117/1.NPh.1.2.025005]

Keywords: functional near infrared spectroscopy; test–retest reliability; infants.

Paper 14014RRR received Feb. 26, 2014; revised manuscript received Aug. 6, 2014; accepted for publication Aug. 11, 2014; published online Sep. 8, 2014.

1 Introduction

The development of noninvasive brain imaging techniques over the last 20 years has led to rapid growth in our understanding of brain function and structure. A major challenge for developmental researchers has been to develop infant-friendly neuroimaging methods. In particular, the development of near infrared spectroscopy (NIRS) for the study of functional brain imaging (fNIRS) in infants has been a welcome addition to the very limited choice of methods currently suitable for the use in awake infants. Over the last decade, fNIRS has become established as an easy-to-use, relatively transportable, and low-cost brain imaging technique. For many years, the primary choice for functional imaging in awake infants has been electroencephalography (EEG), a noninvasive technique with high temporal resolution but relatively poor spatial resolution. A major advantage of fNIRS compared with EEG is that it is less susceptible to data corruption by movement artifacts and offers more highly spatially resolved images of activation allowing the localization of brain responses to specific cortical regions. fNIRS is similar to fMRI in that it can measure the hemodynamic response to neuronal activation. The spatial resolution and depth sensitivity are lower than that of fMRI;[1,2] however this has not prevented the technique from finding widespread use as a neuroimaging tool where other techniques are not practically applicable. Specifically, the use of fNIRS to study functional brain activation in infants is a rapidly growing research area.[3,4] To date, the technique has been used to address developmental topics such as object processing,[5] social communication,[6–8] human action processing,[9,10] and face processing,[11] and it has recently been extended to research on atypical trajectories of brain development, such as in developmental disorders.[12,13]

A recent shift in the use of fNIRS has been toward the study of the infant brain on an individual level.[14–16] This form of analysis is particularly important in prospective longitudinal studies of infants at risk, as it enables the comparison of brain activity with behavioral and demographic data across a variety of measures. Furthermore, the assessment of individual differences in infants' responses is necessary for the discovery of early warning markers in infants at risk for compromised neurodevelopment[17] and consequently for the development of prodromal interventions. However, in order for us to accurately measure individual differences in brain activation, it is essential to first identify the factors influencing reliability and then to quantify their contribution to measurement variability. Hence, reliability is a crucial issue in functional activation measurements, as the ability to detect individual differences will be compromised if the reliability of the method is questionable.

Studies of retest reliability in adults have been conducted with other imaging techniques such as fMRI[18,19] with a wide range of reported values of reliability depending on the number of participants in the study, the number of task runs, and the tasks used to test reliability.[20] Reliability studies have also been conducted with EEG[21–23] in adults, showing strong reliability of imaging measurements. Test–retest studies on fNIRS have been published on adults in muscle[24,25] and brain function[26–30] studies. However, to our knowledge, there are no fNIRS

*Address all correspondence to: A. Blasi, E-mail: anna@blasi.com; S. Lloyd-Fox, E-mail: s.fox@bbk.ac.uk

†These authors contributed equally to this work and their names are ordered alphabetically.

reliability studies published with infants. Comparisons of group fNIRS data across different publications can be difficult because of variations in stimuli, testing designs, probe placements, criteria for data rejection, signal processing, and statistical analysis methods.[31] Longitudinal studies in the same individuals can allow for standardization of some of these sources of variation and therefore provide more appropriate data from which to draw conclusions about the reliability of fNIRS data. Once known, these measures of reliability in young populations will allow us to establish whether the technique provides sufficiently robust measures of individual differences to establish longitudinal associations in human development. Given that the number of published infant fNIRS studies now exceeds 100, it is surprising that test–retest reliability analyses have thus far not been undertaken. However, this may in part be due to the fact that infants can rapidly habituate to repeated stimuli or task demands, and in contrast to adults cannot be asked to attend on demand, making repeated sessions vulnerable to lost trials and poor compliance.[32] Further, infants are capable of remembering events and retain memories of these from very early in life. At 4 to 8 months, they retain the memory of a single task for a few weeks or longer with reminders.[33] Thus, it is safer to increase the retest interval to a few months rather than to a few weeks with a test and retest study with young infants in order to ensure best repeatability of the construct and improve participant compliance with the study. In support of this approach, previous test–retest data from adults shows that stimulus-specific decreases in the cortical response with repeated exposure are evident when a short retest interval is used (3 weeks) but not a long (up to 53 weeks) interval.[29] While there are also limitations of collecting data from more distant test sessions (several months apart), this has the critical advantage of better data quantity and quality at the second test session. Therefore, we investigated the test–retest reliability of measuring hemodynamic brain responses using fNIRS with a cohort of infants who were participating in a longitudinal study over a 9-month period.

The current work aimed to investigate the following questions. First, how replicable are the significant group effects across two data acquisition sessions? This will be assessed by how many fNIRS channels show significant hemodynamic responses during a functional paradigm, and how similar the spatial group maps of activation are across two time points. Second, how replicable are the significant hemodynamic responses within individual infant data across two time points? This will be assessed by measuring the similarity in spatial maps at the individual participant level. And third, how replicable are the measured signal changes (of the hemodynamic time course as a whole) at group and individual levels in repeated sessions? This will be assessed by comparing time courses and variability of the acquired data between the sessions.

2 Materials and Methods

2.1 *Participants*

The data for this analysis was retrospectively selected from a group of infants who were enrolled in a longitudinal fNIRS study in The Gambia.[34] The number of participants recruited for the original study was 42, and the 13 infants included in the study were selected based on availability of valid data for two data acquisition sessions. From the original 42 infants recruited, 18 were excluded from Session 1 due to insufficient number of valid trials according to looking time measures (seven infants), experimenter error (seven infants), or tiredness/fussiness (four infants). Of the remaining 24, one infant died before Session 2 took place, one family moved away from the region, and a further nine participants were excluded from Session 2 due to an insufficient number of valid trials as assessed by looking time (four infants) or tiredness/fussiness (five infants). Session 1 was conducted when the infants were 4 to 8 months of age (175 days $\pm$ 40.19), whereas Session 2 was conducted when the infants were 12 to 16 months of age (432 days $\pm$ 36.56), and the average retest interval was 8.5 months (256.2 $\pm$ 5.4 days).

Participants were identified from the West Kiang Demographic Surveillance System.[35] All infants were born full term (37 to 42 weeks' gestation) and with normal birth weight. Ethical approval was given by the joint Gambia Government/MRC Unit The Gambia Ethics Committee, and written informed consent was obtained from all parents/carers prior to participation.

2.2 *Experimental Procedures*

Details of the experimental design are described in previous publications.[14,34,36] Infants wore custom-built fNIRS headgear consisting of an array over the right hemisphere (see Fig. 1), containing a total of 12 channels (source–detector separations;

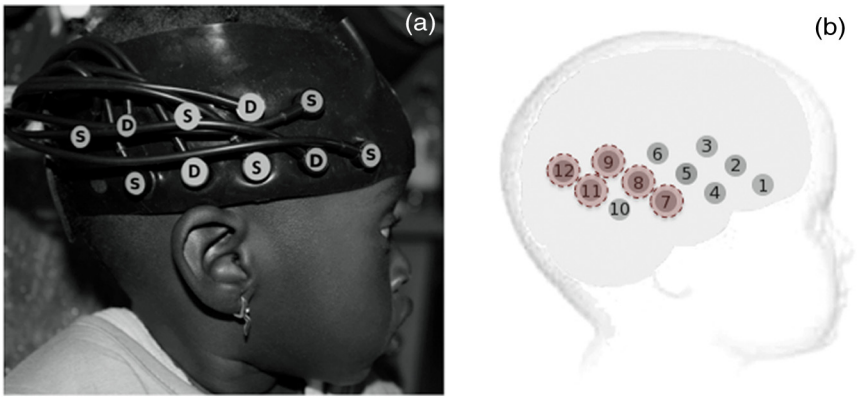

Fig. 1 Near infrared spectroscopy (NIRS) sensor array used in the studies. (a) Participant wearing the headgear, showing the distribution of source (S) and detector (D) optodes. (b) Representation of the location of the channels; highlighted circles indicate channels included in the region of interest (ROI).

2 cm) and were tested with the UCL optical topography[37] system. Note that measurements were restricted to the right hemisphere as (1) our funding only allowed for a restricted number of sources and detectors with respect to the NIRS system used in the UK and (2) we localized the channels to one hemisphere to ensure we could measure the full extent of the temporal lobe. This system uses near-infrared light of two different wavelengths (780 and 850 nm). Before the infants began the study, head measurements were taken to align the headgear with 10 to 20 coordinates.[14] Measurements from this group of infants showed that the average head circumference was 41.3 cm (SD = 1.08) in Session 1 and 44.2 cm (SD = 1.47) in Session 2. The headgear was placed over the right hemisphere with the source optode between channels 4 and 7 centered above the preauricular point (directly over T4 according to the 10-20 system). The angle of the positioned array was guided by the headband, which was placed on the head so that it touched the top of the ear (where the ear joins the head) and lay over the brow line of the infant (through Fp1 and Fp2). According to the head measurements of the 4 to 16 months in the current study, in this position the most anterior optode was positioned approximately over F8. Though the head circumference for this age range is smaller in this Gambian population compared with WHO standards, the relative increase in size between the two age points is similar.

The experimental protocol was identical in both sessions. This experimental design had been successfully used in previous studies in the UK, to investigate responses to auditory and visual social stimuli in typically developing infants and to compare responses with infants at risk for developmental disorders.[12,14] Infants sat on a parent's lap in front of a screen. The parent was instructed to refrain from interacting with the infant during the stimuli presentation unless the infant became fussy or sought their attention. The conditions alternated one after the other, with a period of baseline between each. Three types of conditions [visual-social (silent) V-S, auditory vocal V, auditory non-vocal N-V] were presented in the same order across infants in a repeating loop (V-S, N-V, V, V-S, V, N-V) of trials (single presentation of a condition). For the current work, we focused on one of the three experimental conditions—auditory vocal—which consisted of full-colour, life-size videos of human motion (i.e., "Peek-a-boo") displayed for 9 to 12 s (average 10 s), accompanied by human vocal sounds (i.e., yawning, crying, laughing) for a duration of 8 s. Each trial consisted of four different sounds presented for 0.37 to 2.92 s each, interleaved by short silence periods of 0.16 to 0.24 s. Vocal stimuli were chosen from the Montreal Affective Voices (for more detail, see Ref. 38) and the stimuli of the voice functional localizer.[39] We measured activation during presentation of this experimental condition compared to the baseline condition, which consisted of nonhuman still images (i.e., cars and houses) presented randomly for a pseudorandom duration (1 to 3 s) for 9 to 12 s (average 10 s) with silence. The trials were presented until the infants became bored or fussy as judged by the experimenter who was monitoring their behavior. On average, participants looked for 5.61 experimental auditory-vocal trials in Session 1 and 6.54 in Session 2 (no significant difference between the two sessions).

2.3 *Behavioral Data Processing*

Each session was videorecorded in order to code offline infant behavior and compliance with the study. A researcher unfamiliar with the study's aims carried out behavior coding from these videos. Due to resource limitations at the time of testing, videos were recorded differently at Session 1 and Session 2. As a result, in Session 1, it was possible to synch them with the start and end of the study, but not with the start of each individual trial, as was done in Session 2. During Session 1, the whole session was coded (but without a record of trial) and data was considered valid when the infant watched for >60% session (in addition, the experimenter noted invalid trials online during the study when the infant looked away), whereas in Session 2 in addition to this coding, the trial transitions were also videoed so the session could be coded trial by trial and data considered valid if the infant watched for >60% of each individual trial (as used in previous work[9]). In Session 2, 12 out of 13 of the infants show the same validity coding for session coding versus trial by trial coding [we included 1 infant in Session 2 who had valid data in the experimental condition under consideration (vocal) but not in the other two experimental conditions]. Furthermore, online experimenter coding was highly reliable, with trial by trial experimenter coding matching the video coding in 10 out of 13 of the infants in Session 2, with 1 invalid trial not coded by the experimenter online in three infants. Therefore, we can be confident that the session video recording and experimenter coding in Session 1 were sufficient.

2.4 *fNIRS Data Processing*

Changes in HbO_2 and HHb chromophore concentration (μmol) from baseline to experimental condition were calculated and used as hemodynamic indicators of neural activity.[40] The same differential pathlength factor (DPF) was used across the two age points[41] (DPF = 5.13), as the variability of DPF with age for each wavelength was minimal.

The data was low-pass filtered and divided into blocks that consisted of 4 s of prestimulus onset baseline, followed by the experimental trial and, after that, a whole trial of baseline (9 to 12 s in length). Each block was detrended by fitting a straight line between the average signal value in the prestimulus onset period and the average signal value on the last four seconds at the end of the block, which correspond to the last part of the subsequent baseline trial. The detrending procedure brings the start and end points of each block to zero, so the HbO_2 and HHb values reflect increase or decrease from that reference value.[3] Measurements for each infant were analyzed, and trials, channels, or participant data were rejected from further analysis in a two-step preprocessing protocol: first, by looking time measures, and second, by the quality of the signals as assessed by artifact-detection algorithms (which either excluded the data of whole channels per infant or data from individual trials within a channel, according to the magnitude of the artifact).[3,42] Criteria for channel rejection included: (1) measuring the coefficient of variation (CV) of the signal (channels were excluded if the CV of the attenuation measurement for each wavelength exceeded 10%, possibly due to movement of the arrays and hat) and/or (2) high-frequency noise beyond the limits of physiological effects, where the normalized high-frequency power is greater than 35% of the total power of the signal.[43] For each infant, the channels that survived these rejection criteria were analyzed for trial selection. The trial selection analysis identified sharp changes in the signal caused by sudden movements. This was applied following data conversion from attenuation to concentration data. Trials that contained changes in HbO_2 concentration that exceeded a predefined range (± 3.5 μmol during baseline and ± 8 μmol during the experimental trials where

artifacts in the signal may occur in addition to activation), were removed from the data set. These thresholds were set according to experience with the current array design over the past 8 years. The minimum number of valid experimental trials for each channel was 3. At group level, the grand averaged hemodynamic responses (μmol) of all infants were calculated and the maximum change (or amplitude) in HbO_2 (increase in chromophore concentration) and/or HHb (decrease in chromophore concentration) was assessed during the experimental condition relative to baseline within a time window selected between 8 and 16 s poststimulus onset for each trial. This period of time was selected to include the range of maximum concentration changes observed across infants for HbO_2 and HHb. Two-tailed *t*-tests were used to test the statistical significance of the change. Either a significant increase in HbO_2 concentration or a significant decrease in HHb is commonly accepted as an indicator of cortical activation in infant work.[3] During the channel by channel *t*-tests and subsequent spatial reliability analyses, if HbO_2 and HHb were either to increase or decrease significantly in unison, the signal was considered inconsistent with a hemodynamic response to functional activation[40] and not reported in the analyses (for further discussion of physiological changes reported in infant fNIRS work, see Refs. 3 and 4). To identify these channels, the statistical analyses were reviewed and those channels with an increase or decrease in both chromophores were excluded. For the group level, no channels evidenced this pattern in either session. For the individual level, in Session 1, three participants had one channel excluded and one participant had eight channels excluded from the HbO_2 activation maps; and in Session 2, two participants had channels excluded (one channel and four channels). This exclusion criterion was not applied during the signal reliability analysis. Throughout the text, the terms "significant increase of HbO_2," "significant decrease in HHb," or "significant channel" will be used considering these criteria. To resolve statistical problems of multiple measurement sites for these group analyses, we applied the false discovery rate (FDR) test for multiple comparisons.[44,45] The channels that did not survive the test are highlighted in Table 1 with an asterisk. HbO_2 results were unaffected by FDR correction; however, none of the channels with significant HHb decrease survived the test.

At the single participant level, statistical significance of signal change within each channel was calculated by two-tailed *t*-test during the 8- to 16-s time window identified at group level. This analysis assessed the average hemodynamic change within a 6-s window centered on the observed maximum change per trial. By using the average within this secondary window, we aimed to reduce potential bias of artifacts in the data, as at this level the analysis considered single trial time courses instead of the average of several time courses. Significant activation was then defined using the same criteria as for the group analysis.

2.5 *Alignment Measures of fNIRS Headgear Placement*

As the precision of repositioning the fNIRS arrays may be subject to some error, it was essential that we made precise measures of the position of the fNIRS array on each individual at each data acquisition session. These were then analyzed with an objective alignment system, referenced to external landmarks on the infant's skull (as recommended by Ref. 31), in order to

Table 1 The results from the channel-by-channel *t*-test (two-tailed) analysis for the contrast between the experimental condition and the baseline for Sessions 1 and 2. For each contrast, the results for the significant increase in HbO_2 and/or decrease in HHb concentration are displayed. Significant signal change is highlighted in bold.

	HbO_2		HHb		HbO_2		HHb	
	t	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Ch	Session 1				Session 2			
1	0.59	0.567	−0.10	0.922	1.92	0.079	−0.90	0.384
2	1.59	0.140	−0.91	0.384	**3.01**	**0.011**	0.53	0.609
3	−1.07	0.306	−0.53	0.607	1.56	0.146	0.05	0.958
4	0.43	0.672	**−2.46**	**0.030**[a]	**3.54**	**0.004**	−1.57	0.143
5	**3.30**	**0.006**	−1.09	0.296	**6.21**	**< 0.001**	**−3.07**	**0.010**[a]
6	**3.11**	**0.009**	−1.00	0.337	**5.32**	**< 0.001**	−1.39	0.189
7	**4.27**	**0.001**	−0.23	0.823	**2.57**	**0.024**	−0.09	0.927
8	**6.71**	**< 0.001**	−1.11	0.290	**4.23**	**0.001**	−1.84	0.090
9	**3.90**	**0.002**	−2.03	0.066	**6.45**	**< 0.001**	**−2.35**	**0.036**[a]
10	−0.02	0.981	−0.42	0.684	−0.45	0.660	1.64	0.127
11	**9.40**	**< 0.001**	−0.45	0.664	**3.96**	**0.002**	**−2.56**	**0.025**[a]
12	**4.68**	**0.001**	−0.44	0.668	**3.84**	**0.002**	−0.50	0.624

[a]Channel tests that would not have survived false discovery rate correction for multiple comparisons.

Table 2 Spatial reliability at group level. A_{S1} = number of significant channels at S1; A_{S2} = number of significant channels at S2; $A_{overlap}$ = number of the same channels significant at both sessions; $R_{quantity}$ = an intersession measure of the size of the response (number of significant channels) at both sessions; $R_{overlap}$ = an intersession measure of the spatial overlap of significant channels at both sessions. Results are given for the HbO_2 response for all available channels in the sensor array [whole array (HbO_2)], the HHb response in the sensor array [whole array (HHb)]; both the HbO_2 and HHb response for all available channels in the sensor array [whole array (HbO_2 and HHb)], and the HbO_2 responses from channels within the ROI.

	A_{S1}	A_{S2}	$A_{overlap}$	$R_{quantity}$	$R_{overlap}$
Whole array (HbO_2)	7	9	7	0.875	0.875
Whole array (HHb)	1	3	0	0.5	0
Whole array (HbO_2 and HHb)	8	9	8	0.941	0.941
ROI (HbO_2)	5	5	5	1	1

record error in fNIRS array placement across the two sessions. To investigate the efficacy of headgear placement across sessions, the position of the arrays on the infants was photographed and head measurements were taken. Due to warping on the images, only linear displacement measurements of the center point of the reference optode (the middle optode on the lower row of the array) in relation to displacement in direction x and y were used to quantify error (see Fig. 2). The alignment grid was overlaid on each photograph (as shown in Fig. 2), and the position of the reference optode in relation to the overlaid axis was recorded. The "zero" error position was taken as the position when the center of the reference optode was aligned with the dorsal to ventral y axis (defined by the position of the tragus and the place at which the ear curves up and away from the head; see Fig. 2) and the lower edge of the headband was aligned with the anterior to posterior x axis (defined by the position of the ear when the top of the ear joins to the head and the highest point of the eyebrows on the photo; see Fig. 2). The diameter of the optode is 10 mm. Therefore, using a scaling factor from actual size (of the optode) to the photo image we were

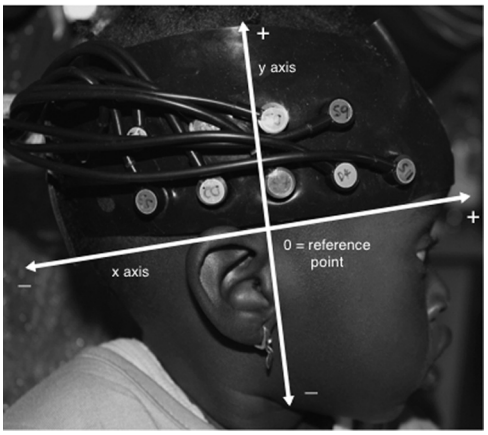

Fig. 2 An example of the reference optode and x and y alignment axes overlaid on the photo of one participant's optode and headgear placement. For this participant, x displacement was 5 mm (1/2 optode diameter) and y displacement was 0 (lower edge of head gear is aligned with the x-axis).

able to calculate how far the optode had deviated from the zero error position for each infant. One limitation of this approach is that errors were only measured in the x and y directions, therefore errors in array rotation were not calculated.

2.6 *Defining a Region of Interest*

As the test–retest analysis was conducted with data from a functional brain activation study, we assessed reliability both over the whole array and within a region of interest (ROI) chosen to assess responses within specific brain regions known to be active during social stimulus paradigms.[12,14,46] We used a standardized scalp surface map of fNIRS channel locators to reliably locate cortical ROI covered by our fNIRS array.[47] This map has been designed to identify ROI within the frontal and temporal lobes for the study of the social brain network in 4- to 7-month olds (with a head circumference ranging from 38 to 45 cm). Though this standardized map may be more applicable for our infant dataset from Session 1 (when they are matched for age), given that the head circumference is smaller in the Gambian cohort compared with the UK infants tested at our lab, we believe that the map can also guide the ROI selection for Session 2. Using the measurement grid of scalp surface channel locations provided by this map,[47] we defined an ROI of channels that were most likely over the superior temporal sulcus region (because the channels were located over either the middle or superior temporal gyri in 75% to 100% of the 55 infants with MRI-fNIRS individual coregistered data[47]). These channels were 7, 8, 9, and 11. We also included channel 12 in the ROI, although the standardized map did not include a channel with the position equivalent to it. However, extrapolation of the position of this channel on the map shows that it would be most likely positioned in the most posterior part of the temporal cortex, but still in the region of the STS. This would be particularly true for the participants at Session 2, when the infants are older. The ROI is shown in Fig. 1.

2.7 *Data Analysis*

Infant compliance with the study was measured using the percentage of time spent looking at the screen (looking time) over the total duration of the session (see Methods). Paired t-tests were used to compare performance between the two sessions.

Spatial reliability of significant HbO_2 and HHb hemodynamic responses were assessed with metrics of size and spatial overlap that have been widely used in both the fNIRS and fMRI literature:[28,29,48]

$$R_{\text{quantity}} = 1 - |A1 - A2|/(A1 + A2),$$

$$R_{\text{overlap}} = 2 \times A_{\text{overlap}}/(A1 + A2),$$

where $A1$ is the number of channels with significant signal change in response to the experimental condition compared with baseline in Session 1, A2 is the number of channels with significant signal change in Session 2, and A_{overlap} is the number of channels with significant signal change in both Sessions 1 and 2. R_{quantity} is a measure of replicability in the size of activation, whereas R_{overlap} is a measure of the replicability of the spatial location of activation. These measures have been used to assess reliability at group and single participant levels in adult fNIRS studies.[26,29]

Further to these significant threshold based analyses, signal reliability of the HbO_2 hemodynamic response was assessed in two additional steps. First, signal reliability was assessed with the Pearson correlation coefficient of the signal hemodynamic time course between the two sessions. At the group level, a Pearson correlation coefficient was conducted on the average hemodynamic time course (averaged across trials, 240 time points according to a time resolution of 10 Hz) to assess the reproducibility of the shape and timing of the signal across channels and participants. At the single participant level, Pearson correlation was calculated using the mean signal change averaged across trials (240 time points) and channels within the ROI per sessions for each participant. Second, for group-level analyses, signal reliability was also calculated with the intraclass correlation coefficient (ICC, one-way random effects[49]). In this work, ICC_{single} at group level is a measure of the ratio of between participants' variance over the total variance and informs about the reproducibility of a single measurement (i.e., for a single participant); and $ICC_{average}$ is a measure of between-session variance over total variance and represents the reproducibility of the mean of repeated measures (i.e., or the replicability of session measurements[50]). ICC values are interpreted as follows: a value of 1.0 would indicate nearly perfect agreement, a value of 0 would indicate there is no agreement, while a negative value should be treated with caution and is thought to be unreliable.[51,52] Reliability measures >0.5 were considered reasonable in previous adult fNIRS and fMRI test–retest studies.[18,29,53]

3 Results

The 13 infants who participated in this study had valid fNIRS data from both sessions (see Methods for measures of validity). First, infant attentiveness and engagement with the study was evaluated, and the common measure of percentage of looking time over total duration of the whole study session (which includes all experimental and baseline conditions) was 77.78% (SD = 9.33) in Session 1 and 87.06% (SD = 8.76) in Session 2. The difference in percentage of looking time between the sessions was not significant (pairwise *t*-test, $t = 1.405$, $p = 0.190$; see Fig. 3). If we focus these analyses on the auditory-vocal experimental condition, the number of trials obtained in Session 2 was on average longer (average number of auditory-vocal trials played in Session 1 = 5.62, SD = 1.12; Session 2 = 7.00, SD = 1.68; $t = 2.920$, $p = 0.013$). However, the average number of trials per participant that achieved looking time criteria (as specified in the Methods) was similar in both sessions (Session 1 = 5.62; Session 2 = 6.54; $t = 1.556$, $p = 0.146$). Second, the artifact-detection algorithms revealed that the data were largely free of artifact. Across the data from both sessions only one infant had two channels excluded; the remaining infants had a complete set of valid channels. Within the data that achieved looking time criteria, the average number of trials excluded per channel within individual infants in Session 1 was 1.23 (SD = 1.74) and in Session 2 was 1.00 (SD = 2.52). Seven infants in Session 1 and 10 infants in Session 2 did not have any trials excluded by the automatic detection of artifacts in any channel; furthermore, 11 infants in both sessions had none or only one trial removed on any of the sessions. Channels 1 and 2 (affecting Session 1 only) were the channels excluded from the participant with channels excluded from the analysis. Overall, automatic artifact detection and exclusion of corrupted trials affected both sessions similarly, as the average percentage of included trials in the analysis after automatic artifact detection for Session 1 = 97.50% (SD = 4.65) and Session 2 = 98.46% (SD = 4.16; $t = 0.521$, $p = 0.612$).

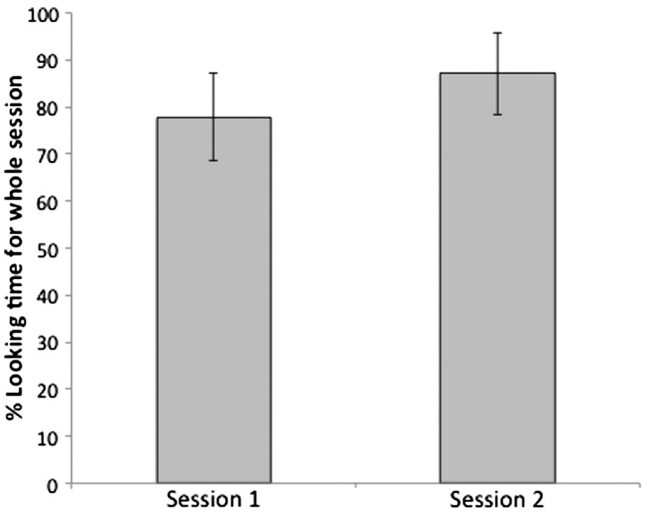

Fig. 3 Average percentage of looking time across participants per Session (mean ± standard deviation). No significant difference was found between sessions in percentage of looking time.

In an initial group analysis, the maximum hemodynamic changes were identified within the time window of interest (see Sec. 2) in response to the experimental condition (auditory-visual social stimuli) versus the baseline (silence with non-social visual stimuli) and analyzed channel-by-channel (*t*-test, two-tailed). This analysis revealed significant increases in HbO_2 and significant decreases in HHb across a wide number of channels (see Fig. 4 and Table 1).

In an initial individual infant analysis, trial-by-trial significant HbO_2 increases in response to the experimental stimulus versus baseline (average responses per trial within the time window of interest; see Sec. 2) were detected in at least one channel, across the whole fNIRS array, in 13 of the infants (100%) at Session 1 and 10 of the 13 infants (77%) at Session 2. Twelve of the 13 infants in Session 1, and all 10 of the infants in Session 2, revealed a significant response in at least two channels. Significant HHb decreases were detected in six out of 13 (46%) infants at Session 1 (four of the six with at least two channels with significant responses), and in 11 of the 13 infants (85%) at Session 2 (six of the 11 with at least two channels with significant responses). Taking into account that the number of significant channels was higher for HbO_2 than HHb across the group of infants (Session 1: average of 3.54 channels with HbO_2 increase, 1.15 channels with HHb decrease; $t = 4.34$, $p < 0.001$; Session 2: average of 4.77 channels with HbO_2 increase, 2.38 channels with HHb decrease, $t = 3.31$, $p = 0.006$), and that all channels with a significant HbO_2 increase passed the FDR test for multiple comparisons, while none of the channels with HHb decrease did, we decided to base our reliability analysis on the most robust measure. Hence, in this work, we mainly focus on HbO_2 changes. However, as it is strongly recommended that both HbO_2 and HHb are included when reporting activation,[3,40] we will also include some measures of HHb reliability where possible (i.e., when activation-related HHb signal changes were observed).

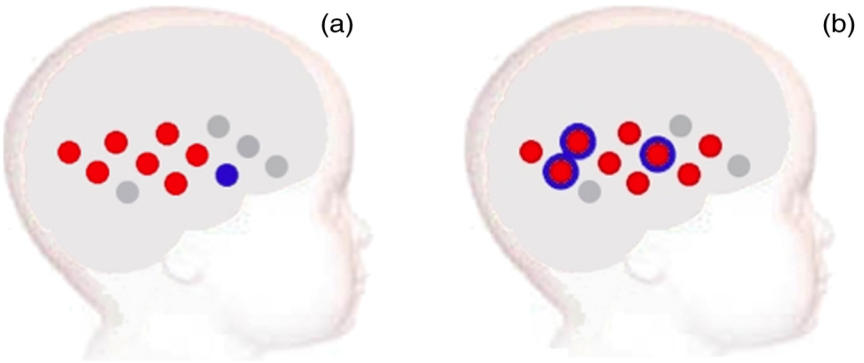

Fig. 4 Significant group results illustrating map wise replicability, (a) Session 1; (b) Session 2. A significant increase in HbO_2 (red), significant decrease in HHb (blue) concentration, or no significant response (gray) is illustrated for each channel.

3.1 *Reliability of fNIRS Headgear Placement*

Placement of the fNIRS array on the individual infant's head did not vary significantly across sessions. In Session 1, in relation to the reference zero position (see Fig. 2; further details in Methods) the reference optode was on average, 2.2 mm (SD = 9.4) more anterior and 1.9 mm (SD = 1.9) more inferior; and in Session 2, was 0.1 mm more anterior (SD = 8.6) and 1.2 mm (SD = 2.3) more superior. The position of the reference optode therefore differed on average by 2.2 mm along the anterior–posterior x-axis (n.s., $t = 0.614$, $p = 0.55$) and 3.1 mm along the superior–inferior y-axis (significant difference, $t = 3.784$, $p = 0.003$). Although the latter difference was significant, 3.1 mm is a comparatively small divergence in relation to the resolution of the fNIRS measures at source–detector separations of 20 mm.

3.2 *Reliability at Group Level*

The reliability of the significant changes in HbO_2 and HHb concentration (in response to the experimental condition versus baseline) across the sessions was first assessed at the group level. Spatial replicability at the group level was high. For HbO_2, seven channels were significant at Session 1, and nine channels were significant at Session 2. For HHb, one channel was significant in Session 1 and three channels in Session 2. Eight out of the nine channels with a significant hemodynamic response (in either HbO_2 or HHb) at Session 2 also showed a significant response at Session 1. Intersession measures of the size ($R_{quantity}$) and the spatial overlap ($R_{overlap}$) of significant channels showed a high degree of replicability in detection of HbO_2 increase ($R_{quantity} = 0.875$; $R_{overlap} = 0.875$); however, in terms of detection of HHb change, spatial replicability was much lower ($R_{quantity} = 0.5$; $R_{overlap} = 0$).

Replicability measures of size and spatial overlap increased further when significant changes in both HbO_2 and HHb were taken into account ($R_{quantity} = 0.941$; $R_{overlap} = 0.941$, see Table 2).

Following this, analyses were undertaken on those channels within the superior temporal sulcus region ROI (defined in Methods). All channels within the ROI showed significant activation on both sessions, therefore, size and spatial overlap measures in this region are 1. For HbO_2, the intersession correlation coefficient of the group hemodynamic time course (averaged across infants and channels within the ROI) was 0.896 (see Fig. 5). Inspection of the correlation coefficients within each channel revealed a high degree of correlation in all channels of the ROI except for channel 7: correlation coefficient in channel 7 = 0.562, whereas the range of correlation coefficients for the remaining channels is 0.831 to 0.968. If the ROI correlation coefficient is reanalyzed with channel 7 excluded, it increases to 0.919. For HHb, the intersession correlation coefficient of the group hemodynamic time course was 0.777. Inspection of the correlation coefficients within each channel revealed a wider range from 0.152 to 0.907, and consistent with the HbO_2 results, the lowest correlation coefficient was found in channel 7.

Signal reliability was measured at the group level for the ROI with ICC measurements calculated using the maximum HbO_2 hemodynamic change (averaged across all ROI channels) per participant, for each session. $ICC_{average}$ represents a measure of intersession reproducibility and ICC_{single} represents a measure of intrasession reproducibility. The ROI analysis revealed an $ICC_{average}$ of 0.461 and an ICC_{single} of 0.299 (see Table 3). At the channel level, ICC was calculated using the average of HbO_2 change per channel within the ROI for each participant and revealed reasonably similar $ICC_{average}$ and ICC_{single} measures in four of the five channels. The output from channel 11 should be treated as unreliable, as a negative ICC value was found.[51,52]

ICC measures were not calculated for HHb given the low number of channels with significant HHb change at group and individual levels.

3.3 *Reliability at Single Participant Level*

Good spatial reliability was found at the single participant level for HbO_2 change. Measures of spatial reliability were calculated using data from the 10 participants with at least one significantly active channel on both sessions, initially considering the whole array. $R_{quantity}$ was 0.66 on average (ranging between 0.22 to 0.92) and was ≥0.5 for eight of these 10 infants. $R_{overlap}$ across the whole array was, on average, 0.45, and individual values ranged between 0.22 and 0.77; $R_{overlap}$ was 0.5 or above in four of the 10 infants. Within the ROI, average size reliability ($R_{quantity}$) was 0.78 (ranging from 0.40 to 1), and in nine out of the 10 participants was above 0.5; $R_{overlap}$ in the ROI was on average 0.55 [ranging between 0 (one infant) and 0.8], and above 0.5 in six out of the 10 infants (see Table 4). Detection of significant HHb change in both Sessions 1 and 2 was achieved in four out of 13 infants (all four infants had significant HbO_2 change in both sessions), and in three of them, the channels with significant HHb change were ROI channels.

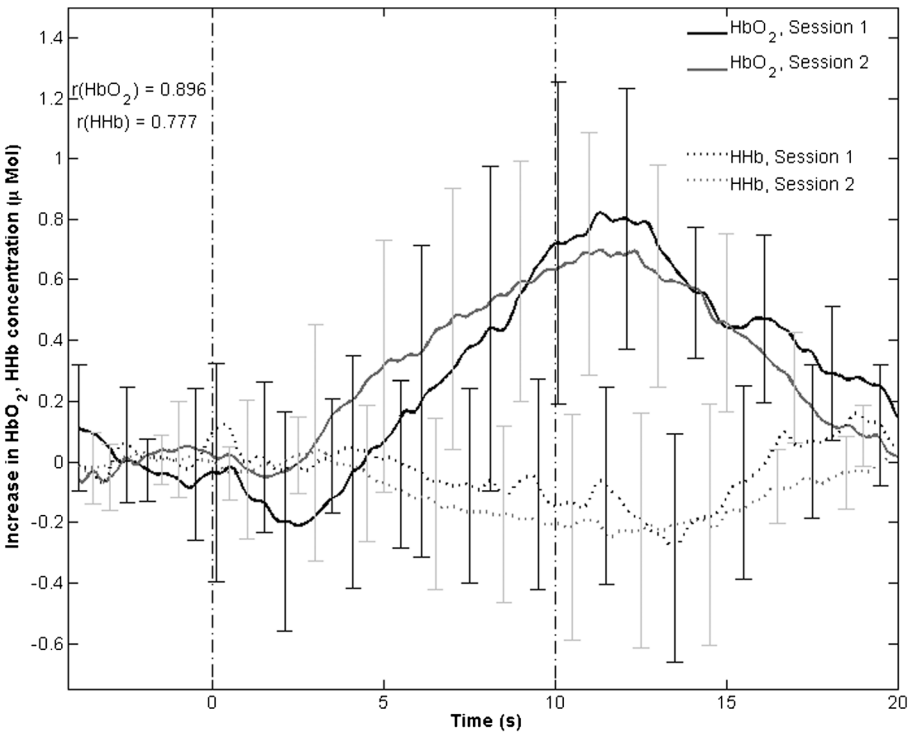

Fig. 5 Mean time course changes in HbO_2 (solid lines) and HHb (dotted lines) across all channels in the ROI and across all infants per session. Dash-point vertical lines indicate the start and end of the task presentation; *r* is the correlation coefficient between the time courses for each chromophore.

Signal reliability of the hemodynamic time course across sessions at the participant level was measured using Pearson correlation coefficient of the signal change averaged across trials and channels within the ROI. This ranged from −0.36 to 0.91 (see Table 5). Six participants showed a correlation above 0.5 (and a further three above 0.4), indicating that their response across the channels within the ROI was consistent across the two sessions. By contrast, four participants revealed negative (or zero) correlation, indicating that their response across the channels within the ROI was not consistent across sessions.

Table 3 Signal reliability at group level for the ROI and across the channels within the ROI. Here, $ICC_{average}$ is a measure of intersession reliability; ICC_{single} is a measure of intrasession reliability (across participants). At ROI-level, ICCs were calculated using the average of the maximum HbO_2 change across all ROI channels per participant. At channel level, ICCs were calculated using the average of the maximum HbO_2 change per participant at each channel.

	ROI	
	$ICC_{average}$	ICC_{single}
ROI	0.461	0.299
Channel 7	0.276	0.160
Channel 8	0.542	0.372
Channel 9	0.415	0.261
Channel 11	−0.338	−0.145
Channel 12	0.633	0.463

4 Discussion

In this work, we have investigated the reliability of using fNIRS to study brain activation over repeated sessions with the same infants, in terms of both reproducibility and similarity in the response. These infants were part of a longitudinal study investigating brain responses to the presentation of auditory-visual social stimuli compared with a silent non-social baseline. Previous research has demonstrated that these types of auditory-visual social paradigms have been associated with activation in the superior temporal sulcus region in early infancy,[9,42] childhood,[54] and adulthood.[55] In the current work, this paradigm was used to assess the reliability of finding similar patterns of significant changes in HbO_2 and HHb over two sessions. The first session was conducted when the infants were 4 to 8 months of age, and the second, 8.5 months later when they were 12 to 16 months old. Clearly, there is the potential for developmental effects to confound our measure of reliability, as the shape, timing, location, or magnitude of the hemodynamic response may change with age during infancy. However, the choice of paradigm used for test–retest in these analyses was designed to minimize these effects, by focusing on a primary functional contrast—auditory stimuli versus silence. Age of participant was not thought to play a significant role, as recent functional imaging studies have revealed that activation patterns to human vocalizations (versus silence) are similar from 3 months of age into adulthood.[46,56,57] Furthermore, though the paradigm included multimodal stimuli, the addition of visual stimuli alongside auditory was not thought to greatly impact the patterns of vocal auditory versus silence activation, as previous studies have found similar results with or without the inclusion of visual input.[14,34,58]

The significant hemodynamic group effects within the ROI were striking in their similarity across test sessions, as was

Table 4 HbO_2 and HHb spatial reliability at single participant level. This includes the 10 participants who had significant HbO_2 and/or HHb responses in at least one channel on both sessions. Results are shown for all channels (whole array) and for the five channels in the ROI.

	HbO_2				HHb			
	Whole array		ROI		Whole array		ROI	
Part. ID	$R_{quantity}$	$R_{overlap}$	$R_{quantity}$	$R_{overlap}$	$R_{quantity}$	$R_{overlap}$	$R_{quantity}$	$R_{overlap}$
001	0.75	0.50	0.80	0.80	0.33	0.33	0.5	0.5
006	0.62	0.62	0.80	0.80	0.40	0.00	0.67	0.00
007	0.86	0.29	0.80	0.40				
015	0.67	0.44	0.67	0.33				
017	0.91	0.55	0.67	0.67				
023	0.92	0.77	1.00	0.75	0.50	0.50		
025	0.50	0.33	1.00	0.67				
026	0.80	0.40	1.00	0.00				
028	0.22	0.22	0.40	0.40	0.60	0.40	0.67	0.00
030	0.36	0.36	0.67	0.67				
Mean	0.66	0.45	0.78	0.55	0.46	0.31	0.61	0.17
$n \geq 0.5$	8	4	9	6	2	1	3	1

Table 5 Signal reliability of the time course at single participant level. Pearson correlation coefficient of mean HbO_2 change within the window of interest (8- to 16-s postexperimental stimulus onset) between Session 1 and Session 2 of the channels within the ROI.

Participant ID	Correlation
001	0.621
003	0.569
006	0.403
007	0.594
015	0.480
017	0.849
023	0.414
025	0.049
026	−0.360
028	0.779
030	−0.246
032	−0.295
035	0.905
$n>0.5$	6

reliability analyzed across the whole fNIRS array. The number of significantly active HbO_2 responses within channels and the spatial overlap of these channels were highly similar across sessions. Therefore, at group level, spatial localization and magnitude of the responses were similar at both test points, making us confident that the fNIRS measurements at group level are robust to potential between-sessions effects such as infant compliance with the study and fNIRS probe positioning on the infant head. These results are in line with long-term fNIRS reliability studies in adults with sessions spread a year apart, though of course the impact of development would be less of an issue there.[29]

As we anticipated, within individual infants, the test–retest results were more variable. Overall, the average individual infant measures of spatial reliability across the whole fNIRS array were at an acceptable level for HbO_2[29] and improved substantially when we focused on our superior temporal sulcus region of interest. For 90% of the infants, the analysis of the number of channels with significant responses revealed $R_{quantity}$ values at or over 0.5. However, there were greater differences when the spatial overlap ($R_{overlap}$) of the significant responses was taken into account, with a wider range across the infants. Therefore, while the magnitude of the response (in terms of number of significant channels) can be seen to be reliable across time, the spatial overlap of the response is more difficult to assess. However, recall that considerable time elapsed between testing sessions and changes in head size, brain morphology, and functional specialization of the response with age may have more impact within individuals than when averaged across a group. For comparison, in adult fMRI studies, mean reliability in spatial overlap at individual level reported values ranging from as low as 0.21 (from a delayed recognition study repeated 1 week apart including six participants) to as high as 0.856 (from a word-generation study

repeated 1 week apart including eight participants, as reviewed by Bennet and Miller[20]). Careful consideration must be taken when comparing changes in signal amplitude across participants or for the same participant across sessions. Location of fNIRS source–detector pairs relative to the site of activation as well as anatomical characteristics such as scalp and skull thickness can have a considerable effect on the amplitude change detected due to partial volume effects.[59] Improvement of single participant measurements can be achieved by using tomographic reconstruction together with anatomical information in models for data analysis. In this work, our reliability results may have been improved had we used an optimal looking time scoring protocol in Session 1 (as we did in Session 2), which would have allowed an accurate exclusion of trials with poor signal (due to lack of attention to the screen) and high noise (with possibly subthreshold movement artifacts).

Our choice to primarily investigate HbO_2 changes was based on its higher signal-to-noise ratio compared to HHb.[60] Furthermore, as the SNR of HHb is lower, the results will be more susceptible to data confounding, such as movement artifact in the data, discrepancies in array placement, and developmental change. While many infant fNIRS studies report significant HbO_2 responses, far fewer report HHb responses, sometimes through choice, but often because they do not find significant responses.[3] This is consistent with the low number of significant group hemodynamic HHb changes seen in the current work. Interestingly, in contrast to the analyses investigating the location and magnitude of the significant hemodynamic changes, the time-course correlation coefficients showed that both the HbO_2 and HHb signal evidenced highly reliable grand-averaged time course data across the two sessions. Future measurements of retest reliability that include HHb reliability within participants should seek to increase the SNR of the signal by increasing participant numbers, designing protocols that elicit strong differential activation in the region of interest, or reducing potential sources of variability in the signals. Furthermore, rather than using fairly basic level statistical tests, more sophisticated analysis techniques such as general linear modeling of the shape of the hemodynamic response may be more sensitive to smaller signal changes and enrich HHb data output in developmental fNIRS studies.

4.1 *Challenges of Gathering Test–Retest Data in Infants*

As we outlined earlier, an aspect of infant development which may impact on the measurement of significant activation at each session is head growth. In other work co-registering individual infants' fNIRS to MRI, we found that age (and not head circumference) is a predictor of changes in fNIRS channel position over underlying anatomy within the range of 4 to 7 months.[47] These findings suggest that growth in head volume (rather than circumference) and changes in the shape and complexity of underlying brain regions may be significant. For example, the shape of the STS may change over age, the depth of the sulci may increase, and therefore the size or shape of the ROI needed to investigate these areas may need to change according to the individual infant's brain morphology. While (1) the co-registered fNIRS-MRI data[48] shows that the location of the channels within our ROI (STG/MTG) is highly consistent across infants and (2) we have designed the ROI to be of sufficient size to accommodate some individual differences in morphology, we acknowledge that in lieu of individual MRI data, we treat the measures of individual reliability with more caution than those of group reliability.

Furthermore, in the current study, we assessed long-term reliability across several months of age. In future work, it would be important to investigate short-term reliability to determine whether the variability in reliability within infants is reduced when age is not a major factor. However, this approach in itself brings considerable challenges, as outlined above.

In conclusion, in this work we demonstrate that (1) spatial mapping and size of activation in infant fNIRS studies has a high degree of reliability and (2) there is strong time course signal reliability within channels of a predefined ROI for group analyses. This work also shows that spatial localization and size of activation in infant populations can be done at the single participant level with an acceptable degree of reliability when a specific region of interest is targeted. Signal reliability results at the single participant level suggest that statistical power may be diminished due to variability of the data at this level. Functional NIRS is, therefore, a highly suitable technique for infant studies, and its reliability at the single participant level can be improved further by adopting strategies that reduce signal variability such as accurate positioning of sensor arrays over regions of interest, regression techniques to examine residual signals at the surface of the head, improving resilience of the sensor arrays to signal artifacts, and accounting further for the changes in brain morphology in the developing brain.

Acknowledgments

We would like to thank the parents and infants who took part in this study as well as the field workers at the MRC Keneba Field Station without whom this work would not have been possible. We thank our collaborators Prof. Andrew Prentice and Dr. Sophie Moore (MRC International Nutrition Group, London School of Hygiene & Tropical Medicine); Dr. Momdou K. Darboe and Dr. Rita Wegmuller (MRC International Nutrition Group, MRC Keneba, MRC Unit, The Gambia); Dr. Maria Papademetriou and Mr. Drew Halliday (Department of Medical Physics and Bionegineering, UCL); and Ms. Katarina Begus (Centre for Brain and Cognitive Development, Birkbeck, University of London). This study was supported by a Bill & Melinda Gates Foundation Phase One Grand Challenges Exploration Grant OPP1061089, core funding MC-A760-5QX00 to the International Nutrition Group by the Medical Research Council UK and the UK Department for International Development (DfID) under the MRC/DfID Concordant agreement, a UK Medical Research Council (G0701484) grant, and a grant from The Simons Foundation (no. SFARI201287 to M. H. J.).

References

1. X. Cui et al., "A quantitative comparison of NIRS and fMRI across multiple cognitive tasks," *Neuroimage* **54**(4), 2808–2821 (2011).
2. G. E. Strangman, Z. Li, and Q. Zhang, "Depth sensitivity and source-detector separations for near infrared spectroscopy based on the colin27 brain template," *PLoS One* **8**(8), e66319 (2013).
3. S. Lloyd-Fox, A. Blasi, and C. E. Elwell, "Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy," *Neurosci. Biobehav. Rev.* **34**(3), 269–284 (2010).
4. J. Gervain et al., "Near-infrared spectroscopy: A report from the McDonnell infant methodology consortium," *Dev. Cognit. Neurosci.* **1**(1), 22–46 (2011).

5. T. Wilcox et al., "Using near-infrared spectroscopy to assess neural activation during object processing in infants," *J. Biomed. Opt.* **10**(1), 1010–1019 (2005).
6. T. Grossmann and M. H. Johnson, "The development of the social brain in infancy," *Eur. J. Neurosci.* **25**(4), 909–919 (2007).
7. T. Grossmann et al., "Early cortical specialization for face-to-face communication in human infants," *Proc. R. Soc. B* **275**(1653), 2803–2811 (2008).
8. Y. Minagawa-Kawai et al., "Prefrontal activation associated with social attachment: facial-emotion recognition in mothers and infants." *Cereb. Cortex* **19**(2), 284–292 (2009).
9. S. Lloyd-Fox et al., "Selective cortical mapping of biological motion processing in young infants," *J. Cognit. Neurosci.* **23**(9), 2521–2532 (2011).
10. H. Ichikawa et al., "Infant brain activity while viewing facial movement of point-light displays as measured by near-infrared spectroscopy (NIRS)" *Neurosci. Lett.* **482**(2), 90–94 (2010).
11. Y. Otsuka et al., "Neural activation to upright and inverted aces in infants measured by near infrared spectroscopy," *NeuroImage* **34**(1), 399–406 (2007).
12. S. Lloyd-Fox et al., "Reduced neural sensitivity to social stimuli in infants at risk for autism," *Proc. R. Soc. B* **280**(1758), 20123026 (2013).
13. S. E. Fox et al., "Neural processing of facial identity and emotion in infants at high-risk for autism spectrum disorders," *Front. Human Neurosci.* **7**, 89 (2013).
14. S. Lloyd-Fox et al., "The emergence of cerebral specialisation for the human voice over the first months of life," *Social Neurosci.* **7**(3), 317–330 (2012).
15. S. Lloyd-Fox et al., "Cortical Activation to Action Perception is associated with action production abilities in young infants," *Cereb. Cortex* (2013).
16. T. Wilcox et al., "The effect of color priming on infant brain and behavior," *NeuroImage* **85**(1), 302–313 (2014).
17. M. Elsabbagh and M. H. Johnson "Getting answers from babies about autism," *Trends Cognit. Sci.* **14**(2), 81–87 (2010).
18. A. R. Aron, M. A. Gluck, and R. A. Poldrack, "Long-term test–retest reliability of functional MRI in a classification learning task," *Neuroimage* **29**(3), 1000–1006 (2006).
19. K. Wagner et al., "The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task," *NeuroImage* **28**(1), 122–131 (2005).
20. C. M. Bennett and M. B. Miller, "How reliable are the results from functional magnetic resonance imaging?," *Ann. N. Y. Acad. Sci.* **1191**, 133–155 (2010).
21. A. J. Fallgatter et al., "Test-retest reliability of electrophysiological parameters related to cognitive motor control," *Clin. Neurophysiol.* **112**(1), 198–204 (2001).
22. K. B. Walhovd and A. M. Fjell "One-year test–retest reliability of auditory ERPs in young and old adults," *Int. J. Psychophysiol.* **46**(1), 29–40 (2002).
23. S. M. Cassidy, I. Robertson, and R. O'Connell, "Retest reliability of event-related potentials: Evidence from a variety of paradigms," *Psychiphysiology* **49**(5), 561–568 (2012).
24. B. Celie et al., "Reliability of near infrared spectroscopy (NIRS) for measuring forearm oxygenation during incremental handgrip exercise," *Eur. J. Appl. Physiol.* **112**(6), 2369–2374 (2012).
25. A. G. Crenshaw et al., "Reliability of near-infrared spectroscopy for measuring forearm and shoulder oxygenation in healthy males and females," *Eur. J. Appl. Physiol.* **112**(7), 2703–2715 (2012).
26. A. Watanabe et al., "Cerebrovascular response to cognitive tasks and hyperventilation measured by multi-channel near-infrared spectroscopy," *J. Neuropsychiatry Clin. Neurosci.* **15**(4), 442–449 (2003).
27. G. Strangman et al., "Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: test-retest reliability and review of the literature," *Archiv. Phys. Med. Rehabil.* **87**(12), S12–S19 (2006).
28. M. M. Plichta et al., "Event-related functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?," *NeuroImage* **31**(1), 116–124 (2006).
29. M. Schecklmann et al., "Functional near-infrared spectroscopy: A long-term reliable tool for measuring brain activity during verbal fluency," *NeuroImage* **43**(1), 147–155 (2008).
30. H. Zhang et al., "Test–retest assessment of independent component analysis-derived resting-state functional connectivity based on functional near-infrared spectroscopy," **55**(2), 607–615 (2011).
31. R. N. Aslin, "Questioning the questions that have been asked about the infant brain using near-infrared spectroscopy," *Cognit. Neuropsychol.* **29**(1–2), 1–2 (2012).
32. C. Rovee-Collier and R. Barr, "Infant learning and memory," Chapter 8 in *Blackwell Handbook of Infant Development*, 2nd ed., J. G. Bremner and T. Wachs, Eds., pp. 271–294, Wiley-Blackwell, Chichester (2010).
33. C. Rovee-Collier, "The development of infant memory," *Curr. Dir. Psychol. Sci.* **8**(3), 80–85 (1999).
34. S. Lloyd-Fox et al., "Functional near infrared spectroscopy (fNIRS) to assess cognitive function in infants in rural Africa," (2014).
35. Medical Research Council (MRC), "The West Kiang Demographic Surveillance System," http://www.ing.mrc.ac.uk/research_areas/west_kiang_dss.aspx (2010).
36. M. D. Papademetriou et al., "Cortical mapping of 3D optical topography in infants," *Adv. Exp. Med. Biol.* **789**, 455–461 (2013).
37. N. Everdell et al., "A frequency multiplexed near-infrared topography system for imaging functional activation in the brain," *Rev. Sci. Instrum.* **76**, 1–5 (2005).
38. P. Belin, "The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing," *Behav. Res. Methods* **40**(2), 531–539 (2008).
39. P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "Montreal Affective Voices (MAV), Audio Collection," http://vnl.psy.gla.ac.uk/resources_main.php (2008).
40. H. Obrig and A. Villringer, "Beyond the visible—imaging the human brain with light," *J. Cereb. Blood Flow Metab.* **23**(1), 1–8 (2003).
41. A. Duncan et al., "Optical pathlength measurements on adult head, calf and forearm and the head of the newborn-infant using phase-resolved optical spectroscopy," *Phys. Med. Biol.* **40**(2), 295–304 (1995).
42. S. Lloyd-Fox et al., "Social Perception in Infancy: A near infrared spectroscopy study," *Child Dev.* **80**(4), 986–999 (2009).
43. J. T. Kirjavainen et al., "The balance of the autonomic nervous system is normal in colicky infants," *Acta Paediatr.* **90**(3), 250–254 (2001).
44. Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann. Stat.* **29**(4), 1165–1188 (2001).
45. A. K. Singh and I. Dan, "Exploring the false discovery rate in multichannel NIRS," *NeuroImage* **33**(2), 542–549 (2006);
46. A. Blasi et al. "Early specialization for voice and emotion processing in the infant brain," *Curr. Biol.* **21**(14), 1–5 (2011).
47. S. Lloyd-Fox et al., "Corregistering functional near-infrared spectroscopy with underlying cortical areas in infants," *Neurophotonics* **1**(2) (2014).
48. S. A. Rombouts et al., "Test–retest analysis with functional MR of the activated area in the human visual cortex," *Am. J. Neuroradiol.* **18**(7), 1317–1322 (1997).
49. P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychol. Bull.* **86**(2), 420–428 (1979).
50. T. Johnstone et al., "Stability of amygdala BOLD response to fearful faces over multiple scan sessions," *NeuroImage* **25**(4), 1112–1123 (2005).
51. R. Muller and P. Buttner, "A critical discussion of intraclass correlation coefficients," *Stat. Med.* **13**(23–24), 2465–2476 (1994).
52. B. Giraudeau, "Negative values of the intraclass correlation coefficient are not theoretically possible," *Clin. Epldemiol.* **49**(10), 1205–1206 (1996).
53. D.S. Manoach et al., "Test-retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects," *Am. J. Psychiatry* **158**(6), 955–958 (2001).
54. K.A. Pelphrey and E.J. Carter, "Charting the typical and atypical development of the social brain," *Dev. Psychopathol.* **20**(4), 1081–1102 (2008).
55. K. A. Pelphrey et al., "Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements," *Cereb. Cortex* **15**(12), 1866–1876 (2005).
56. P. Belin et al., "Voice-selective areas in human auditory cortex," *Nature* **403**(6767), 309–312 (2000).
57. G. Dehaene-Lambertz, S. Dehaene, and L. Hertz-Pannier, "Functional neuroimaging of speech perception in infants," *Science* **298**(5600), 2013–2015 (2002).
58. Y. Minagawa-Kawai et al., "Optical brain imaging reveals general auditory and language-specific processing in early infant development," *Cereb. Cortex* **21**(2), 254–261 (2011).

59. T. J. Huppert et al., "Quantitative spatial comparison of diffuse optical imaging with blood oxygen level-dependent and arterial spin labeling-based functional magnetic resonance imaging," *J. Biomed. Opt.* **11**(6), 064018 (2006).
60. G. Strangmann et al., "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," *NeuroImage* **17**(2), 719–731 (2002).

Anna Blasi is a research fellow at the Centre for Brain and Cognitive Development, Birkbeck, University of London. Her research interests are centered on functional aspects of human physiology. Her research career started with models of the cardiovascular system and the effects of disease. Through her work at UCL, KCL, and Birkbeck, her research interests have shifted toward the use of functional imaging (fNIRS, fMRI) to study brain function and neurocognitive development in early infancy.

Sarah Lloyd-Fox is a research fellow at the Centre for Brain and Cognitive Development, Birkbeck, University of London. Her work focuses on the use of fNIRS to investigate the developing brain in infancy. Her research projects focus on investigating social cognition, human action perception, autism, and most recently, the application of fNIRS in novel settings, such as resource-poor countries to be able to study the effects of compromised development, such as undernutrition.

Mark H. Johnson is a Medical Research Council scientific programme leader and Director of the Centre for Brain & Cognitive Development, Birkbeck (University of London). He is also a Fellow of the British Academy and the Cognitive Science Society. He has published over 250 papers and 10 books on brain and cognitive development in human infants and other species. His laboratory currently focuses on typical and atypical functional brain development during infancy and childhood.

Clare Elwell is a professor of medical physics in the Department of Medical Physics and Bioengineering at UCL. She leads the near infrared spectroscopy (NIRS) research group developing novel optical systems for monitoring and imaging the human body and brain. Her research projects include studies of autism, acute brain injury, sports performance, migraine, malaria, depression, and, most recently, the effects of malnutrition on brain development with the first infant functional brain imaging study in Africa.