Neurophotonics

Neurophotonics.SPIEDigitalLibrary.org

Motion correction for infant functional near-infrared spectroscopy with an application to live interaction data

Hannah F. Behrendt
Christine Firk
Charles A. Nelson, III
Katherine L. Perdue

SPIE.

Hannah F. Behrendt, Christine Firk, Charles A. Nelson, , IIIKatherine L. Perdue, "Motion correction for infant functional near-infrared spectroscopy with an application to live interaction data," *Neurophoton.* **5**(1), 015004 (2018), doi: 10.1117/1.NPh.5.1.015004.

Motion correction for infant functional near-infrared spectroscopy with an application to live interaction data

Hannah F. Behrendt,[a,b] Christine Firk,[b] Charles A. Nelson III,[a,c,d] and Katherine L. Perdue[a,c,*]
[a]Boston Children's Hospital, Laboratories of Cognitive Neuroscience, Boston, Massachusetts, United States
[b]University Hospital RWTH Aachen, Child Neuropsychology Section, Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Aachen, Germany
[c]Harvard Medical School, Boston, Massachusetts, United States
[d]Harvard Graduate School of Education, Cambridge, Massachusetts, United States

Abstract. Correcting for motion is an important consideration in infant functional near-infrared spectroscopy studies. We tested the performance of conventional motion correction methods and compared probe motion and data quality metrics for data collected at different infant ages (5, 7, and 12 months) and during different methods of stimulus presentation (video versus live). While 5-month-olds had slower maximum head speed than 7- or 12-month-olds, data quality metrics and hemodynamic response recovery errors were similar across ages. Data quality was also similar between video and live stimulus presentation. Motion correction algorithms, such as wavelet filtering and targeted principal component analysis, performed well for infant data using infant-specific parameters, and parameters may be used without fine-tuning for infant age or method of stimulus presentation. We recommend using wavelet filtering with $iqr = 0.5$; however, a range of parameters seemed acceptable. We do not recommend using trial rejection alone, because it did not improve hemodynamic response recovery as compared to no correction at all. Data quality metrics calculated from uncorrected data were associated with hemodynamic response recovery error, indicating that full simulation studies may not be necessary to assess motion correction performance.
[DOI: 10.1117/1.NPh.5.1.015004]

Keywords: functional near-infrared spectroscopy; motion correction; simulation; infant; live stimulus presentation.

Paper 17116R received Sep. 13, 2017; accepted for publication Jan. 12, 2018; published online Feb. 13, 2018.

1 Introduction

Functional near-infrared spectroscopy (fNIRS) is an important tool in elucidating the neural underpinnings of many aspects of perceptual and cognitive development in the first years of life (for recent reviews, see Refs. 1–3). However, the use of fNIRS with awake, behaving infants is complicated by the fact that infants are generally allowed to move freely during experimental sessions to facilitate compliance with the experiment. Consequently, infant fNIRS data are often collected with a high degree of participant movement during recordings. Additional constraints such as low trial numbers due to short infant attention spans highlight the need for good motion correction for this important application of fNIRS.

Most work in motion correction for fNIRS data using a simulation approach has focused on adults[4–7] or neonates[8] who are often sleeping during the recording. However, it is unclear if motion correction algorithms developed for those relatively high signal-to-noise ratio cases are applicable to data collected from infants who are awake. While there is no *a priori* reason that the mechanics of spike motion artifacts would differ between adults and infants, the fact that infants cannot be instructed to avoid moving and, therefore, have a high level of artifact in the data suggests that even if the same motion correction methods can be applied to infant data as to adult data, there may need to be some tuning of the parameters for the methods to perform well. One recent study tested motion correction for real fNIRS data collected from children ages 6 to 12 years during a language task.[9] Hu and colleagues examined the heterogeneous nature of artifacts in the data and showed that fNIRS data of child participants contain generally more movement artifacts than fNIRS data of adult participants. They also showed that stacking motion correction algorithms (i.e., using moving average and wavelet filtering) was most effective for child data.[9]

Aside from general differences between movement artifacts in infant, child, and adult fNIRS data, motor development in infants and young children may impact the quantity and scale as well as the shape of motion artifacts in participant groups of different infant ages. In this study, we, therefore, combined quantifying motion using an accelerometer during video stimulus presentation with a simulation approach to test the utility of conventional motion correction methods on infant fNIRS data. We focused on wavelet filtering and targeted principal component analysis (tPCA) motion correction due to their good performance in previous simulation studies.[4–8] Wavelet motion correction is particularly appealing because it can often preserve all trials in a session, which can be critical for infant data.[4,6,8] On the other hand, tPCA allows us to examine only periods of data that have been identified as coinciding with motion artifacts, which is advantageous because it may not incidentally remove

*Address all correspondence to: Katherine L. Perdue, E-mail: katherine.perdue@childrens.harvard.edu

2329-423X/2018/$25.00 © 2018 SPIE

the desired functional signal from the data resulting in a more accurate estimation of the true hemodynamic response function (HRF).[7] We also tested the impact of using more than one motion correction algorithm, in this case tPCA and wavelet filtering, on infant data.

Moreover, given that NIRS is less susceptible to movement artifacts than other imaging tools, it may be possible to use fNIRS to measure the infant brain while reacting to "live" stimuli. For example, live stimulus presentation designs could allow investigation of the developing brain in a more naturalistic and noisy task setting with experimentally standardized, directly observable face-to-face interaction sequences.[10] Live application has been shown to be successful in adult fNIRS studies (e.g., Refs. 11–13) and is a very promising and new development in the field that could contribute to our essential understanding of the early developing brain in its natural context,[10] such as early mother–child interaction. However, it remains to be tested whether live stimuli yield significantly more motion artifacts than video stimuli in infant data and whether motion correction well suited for infant data collected during standard conditions, e.g., video stimulus presentation, are applicable to infant data collected during live stimulus presentation.

This study had three main aims: first, we quantified fNIRS probe motion and data quality metrics during video stimulus presentation for infants over the first year of life. Our participants included infants at 5, 7, and 12 months of age ($n = 20$ per age group). Second, we used infant fNIRS data collected during video stimulus presentation at all three ages combined with simulated hemodynamic responses to test the ability of conventional motion correction methods to eliminate movement artifacts and compared performance of infant-specific tuning parameters. The motion correction methods tested are basic trial rejection, wavelet filtering, and tPCA as well as stacking of motion correction algorithms, i.e., consecutive application of first tPCA and then wavelet filtering. Third, we applied these motion correction methods to infant fNIRS data collected during live stimulus presentation in a separate sample of 6- to 8-month-old infants ($n = 10$). Here, we compared live stimulus presentation versus video presentation in age-similar samples to, first, quantify and compare data quality metrics and, thereafter, test the performance of motion correction algorithms and parameters on live interaction data.

2 Methods

2.1 *Infant fNIRS Datasets*

2.1.1 *Video stimulus presentation and data collection*

Separate groups of 5-month-old ($N = 20$; 9 females, mean age: 152.00 ± 4.62 days), 7-month-old ($N = 20$; 9 females, mean age 213.15 ± 4.20 days), and 12-month-old ($N = 20$; 7 females, mean age 366.80 ± 4.01 days) participants were randomly selected for analysis. The present sample is a subset of an infant sample recruited from an Institutional Review Board (IRB)-approved registry of local births set up by the Laboratories of Cognitive Neuroscience at Boston Children's Hospital/Harvard Medical School (Boston, Massachusetts) to participate in a longitudinal study on emotion processing (for example, Ref. 14 or results presented elsewhere). Infants were typically developing, born full term, with no known prenatal or perinatal complications.

A Hitachi ETG-4000 continuous-wave fNIRS system with wavelengths of 695 and 830 nm was used to collect the hemodynamic responses. Hat design included 46 channels, which spanned over the frontal and bilateral temporal cortices. The probe layout and hat design are shown in Fig. 1(a). Source–detector distances were ~3 cm, and the system sampling frequency was 10 Hz. All infants, regardless of age, were measured using the same hat. A triaxial accelerometer (TSD109C1, BIOPAC Systems Inc., Goleta, California) was attached to the frontal panel [displayed in green in Fig. 1(a)] and used to quantify probe/head motion over the course of the experiment. Accelerometer sampling frequency was 1000 Hz.

During the experiment, infants were seated on a parent's lap while they viewed the presentation screen ~63 cm away. fNIRS and accelerometer data were recorded during a 2-min video presentation. The video was a standard video of infant toys. The Boston Children's Hospital IRB approved the experimental study protocol, and the infant's parent gave informed consent before starting the study session. The infant's parent was paid $20 (USD), parking expenses were refunded, and infants received a toy after their study participation.

2.1.2 *Live stimulus presentation and data collection*

Participants tested were 6 to 8 months old ($N = 10$; 6 females, mean age 230.20 ± 17.51 days). Data from additional, $N = 2$, infants were collected but excluded for the following reasons: infant did not sit through at least three trials per stimulus category ($N = 1$) or technical malfunction during the experiment ($N = 1$). For this separate study, infants were recruited by the University Hospital RWTH Aachen, Germany, in cooperation with maternity clinics in the catchment area of Aachen, Germany, to participate in a study on early mother–child interaction (for example, Ref. 15 or results presented elsewhere). Infants were typically developing, born full term, with no known prenatal or perinatal complications.

A Hitachi ETG-4000 continuous-wave fNIRS system with wavelengths of 695 and 830 nm was used to collect the hemodynamic responses. Hat design included 22 channels that spanned over the frontal cortex. The probe layout and hat design are shown in Fig. 1(b). Source–detector distances were ~3 cm, and the system sampling frequency was 10 Hz. All infants, regardless of age, were measured using the same hat design (EasyCap for fNIRS, EASYCAP GmbH, Germany) but with different hat sizes. No accelerometer data were collected in this study.

During the experiment, infants were seated in a baby's high chair across from the mother and female stranger, respectively, while they viewed the live stimulus presentation ~60 cm away. fNIRS data were recorded during a live face-to-face interaction with mother versus stranger. The live stimuli consisted of 16-s blocks of stressful versus nonstressful face-to-face interaction sequences (still-face versus happy-face expressed by mother versus stranger) each followed by an 8-s rest block (no live stimulus presentation). Each stimulus block was repeated three times. Infants first interacted with their mother and then with the stranger. The total duration of live stimulus presentation was ~6 min. The IRB of the University Hospital RWTH Aachen approved the experimental study protocol, and the infant's parents gave informed consent before starting the study session. The infant's mother was paid 20 EUR, travel and parking expenses were refunded, and infants received a toy after their study participation.

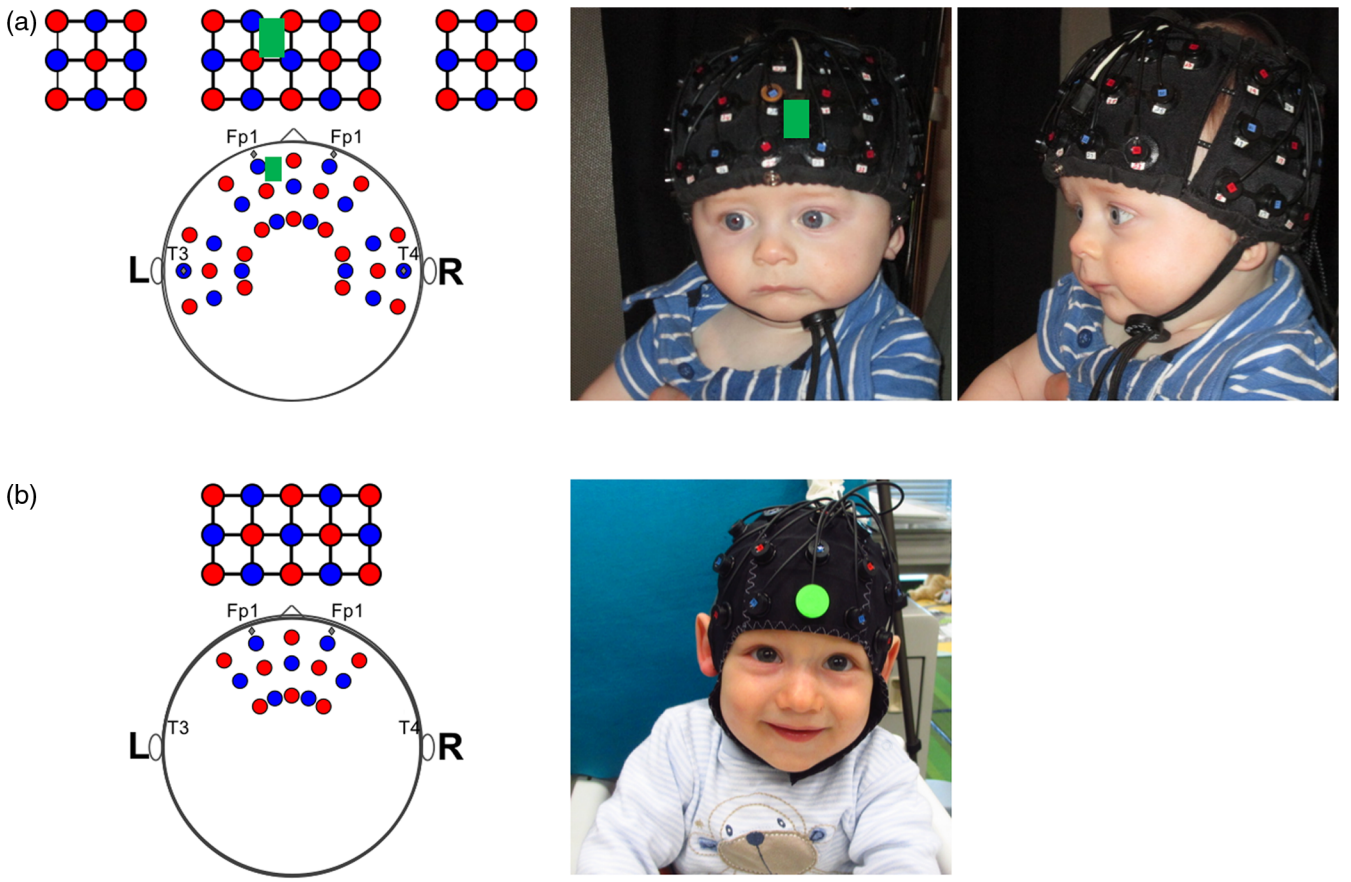

Fig. 1 Probe layout and hat design for (a) infant fNIRS data collected during video stimulus presentation (frontal and temporal panels with a total of 46 channels, sources are displayed in red, and detectors in blue), and accelerometer attached to the frontal panel (displayed in green) and (b) infant fNIRS data collected during live stimulus presentation (frontal panel only with a total of 22 channels, sources are displayed in red, and detectors in blue). In this separate study, no accelerometer data were collected.

2.2 *Motion Correction Methods*

We tested the performance of the following four motion correction methods: basic trial rejection, wavelet filtering, tPCA as well as "stacking" of tPCA and wavelet filtering, and compared their performance against no-motion correction. Most motion correction algorithms require selection of an algorithm-specific tuning parameter that can significantly affect the performance of the motion correction. Thus, we also tested what tuning parameters are well suited for infant data and if parameters need to be adjusted for infant age.

2.2.1 *Basic trial rejection*

Trial rejection is one of the most commonly used motion correction methods for functional fNIRS data,[6] especially for infant fNIRS research, but to the best of our knowledge it has not been rigorously tested in pediatric populations. Stimulus trials that coincide with motion artifacts (identified channel-by-channel and flagged across channels) are removed and not included in the calculation of the subject average. In a first step, motion artifacts are identified based on temporal features of the signal timecourse within each channel using automated artifact detection. Here, usually a specific set of parameter thresholds is applied to identify signal change indicative of spike motion artifacts: for example (i) changes in absolute amplitude of the signal (ampThresh) or (ii) changes relative to the standard deviation of the signal timecourse (stdThresh) within a predetermined time period (tMotion) and time window (±tMask), which is masked around the identified motion artifact. In a second step, stimulus trials are removed. The performance strongly depends on the number of motion artifacts identified and the size of the data sample, i.e., the total number of stimulus trials collected per subject. In this work, we used a lenient trial rejection threshold to preserve as much data as possible due to the low trial numbers typically present in infant data.

2.2.2 *Wavelet filtering*

The wavelet filtering described by Molavi and Dumont[8] employs the Wavelab 850 toolbox for MATLAB®. Wavelet filtering is a motion correction method that is applied to the signal timecourse within each channel independently. In this work, we evaluated the performance of the wavelet algorithm and tested the sensitivity of the tuning parameter *iqr*, which relates to the interquartile range of the wavelet coefficient distribution. It is commonly used to detect and remove motion artifacts, and increasing the *iqr* will delete fewer motion artifacts. Molavi and Dumont[8] set the probability threshold α to 0.1, $\alpha = 0.1$ is equivalent to $iqr = 1.5$. The performance of the wavelet filtering depends strongly on the amount of variance in the data.

2.2.3 *Targeted principal component analysis*

The tPCA described by Yücel et al.[7] is a motion correction method that is applied to the signal timecourse across all channels. In contrast to standard PCA,[16] tPCA employs PCA only on segments of data identified as coinciding with motion artifacts (which are identified by using automated artifact detection in a prior step). The tuning parameter commonly used to detect and remove motion artifacts is (i) the amount of variance (nSV, in %) or (ii) the number of components (N_{PCA}), decreasing it will delete fewer motion artifacts. In this work, we evaluated the performance of the tPCA algorithm as described by Yücel et al.[7] and tested the sensitivity of the parameter stdThresh, which is one of the parameter thresholds used to identify motion artifacts in the data. The performance of the tPCA depends strongly on how motion artifacts are determined based on changes in shape and/or frequency of the signal timecourse. Yücel et al.[7] used the following parameter thresholds: tMotion = 0.5 s, tMask = 1 s, stdThresh = 20, ampThresh = 5, and nSV = 0.97 (97%), and the tPCA is repeated up to a maximum of three iterations (as further iterations did not improve the results; see Ref. 7).

2.2.4 *Data processing streams*

To address the three main aims of this study, we assessed motion correction methods applied to (a) the fNIRS timecourse data for all age groups in the video and live conditions and (b) the fNIRS timecourse data from the video conditions (all age groups) combined with simulated hemodynamic responses. These two complementary approaches allowed us to test how the motion correction methods affect the data without making assumptions about the shape of the hemodynamic response and to quantify how accurately the methods can recover a known hemodynamic response. The processing steps for applying each motion correction method on infant fNIRS datasets (video: 5 months, 7 months, 12 months, and live) are shown in Fig. 2. Data processing was performed using the Homer2 package[17] in MATLAB® (MathWorks, Natick, Massachusetts).

Datasets collected during video presentation included a total of 120 s of fNIRS timecourse data. We, therefore, matched the timecourse length of video presentation and live interaction data; specifically, we used 120 s of live interaction data that included mother–infant stressful versus nonstressful interaction sequences only.

Hemodynamic responses were simulated using a gamma function in MATLAB® (MathWorks, Natick, Massachusetts). Simulated hemodynamic responses were added to the optical density (OD) timecourses of each channel in each dataset. Similar to a standard infant fNIRS block design, we constructed the simulated HRF (see Fig. 7) with a time-to-peak of 6 s and a duration of 16 s, which when converted in to chromophore concentration change [using the modified Beer–Lambert law[18,19] with a differential pathlength factor (DPF) = 5[20]] showed a maximum increase in oxyhemoglobin (oxyHb) of 1 μm and a maximum decrease in deoxyhemoglobin (deoxyHb) of 0.4 μm. This change in chromophore concentration matches average hemodynamic responses that we see in infant fNIRS data collected during a standard experimental block design (for example, Ref. 14). In this simulation study, five hemodynamic responses were added at random intervals to the 120-s infant fNIRS recordings collected during video stimulus presentation in all 46 channels with an intertrial interval (ITI) between 2 and 10 s. We generated the same set of random interstimulus

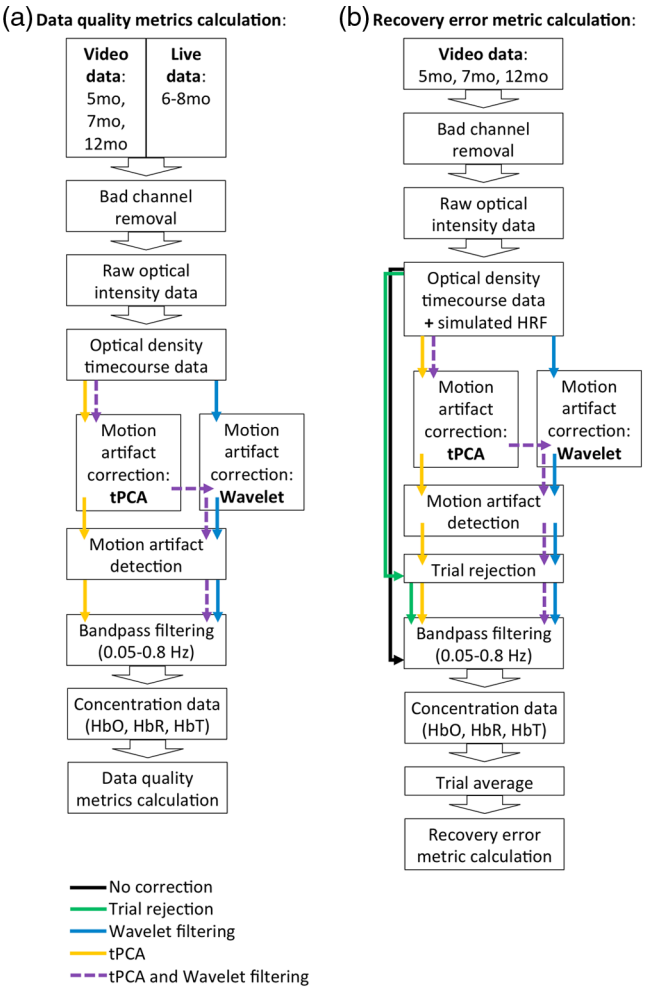

Fig. 2 Data processing streams for (a) calculation of data quality metrics computed for fNIRS timecourse data for all age groups in the video and live conditions and (b) calculation of hemodynamic response recovery error metric computed for fNIRS timecourse data from the video conditions (all age groups) combined with simulated hemodynamic responses.

intervals for all channels per subject and a different set of random ITIs for every subject. Hemodynamic responses were 16-s long and spaces between responses were 2- to 10-s long.

The processing steps for the simulation study are described in detail as follows and are shown in Fig. 2(b); the processing steps for quantifying data quality metrics are shown in Fig. 2(a).

Before conversion into change in OD, channels in each dataset were excluded for artifact, if the magnitude of the raw optical intensity signal was >98% or <2% of the total range for longer than 5 s during the recording as this usually indicated problems, such as low light levels or railing, that are not fixable with motion correction methods. Each dataset was then passed through five different processing streams (no correction, trial rejection, wavelet, tPCA, and stacking). In the first processing stream, we did not apply motion correction before calculating the data quality metrics or HRF recovery error metric. In the second processing stream, we applied basic trial rejection only, as described already, using the motion artifact detection implemented in Homer2 as hmrMotionArtifact with the following set of parameter thresholds: tMotion = 1 s, tMask = 1 s, stdThresh = 25, and ampThresh = 1 to identify motion artifacts and remove affected trials before calculation of the HRF recovery error metric.

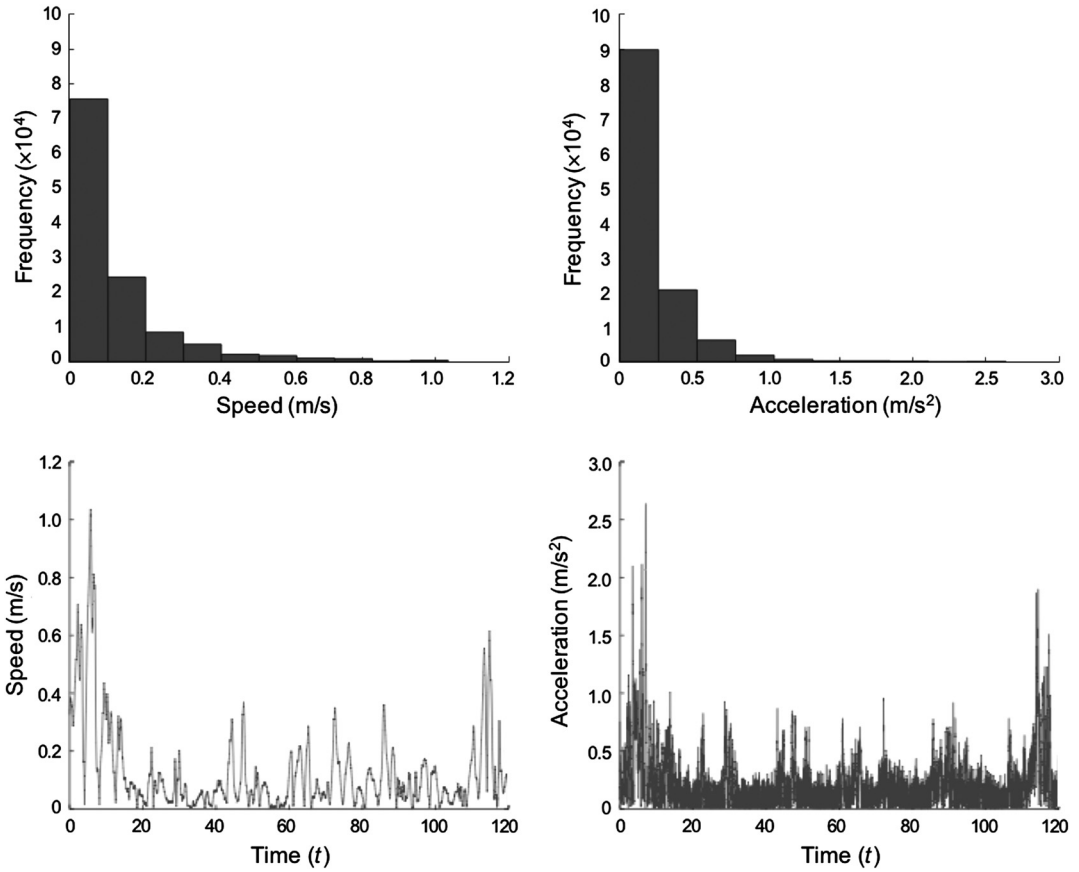

Fig. 3 Motion trace for an example 7-month-old participant.

In the third and fourth processing streams, we implemented the two motion correction algorithms: wavelet filtering and tPCA, respectively, as described already. Wavelet filtering (implemented in Homer2 as hmrMotionCorrectWavelet) and tPCA (implemented in Homer2 as hmrMotionCorrectPCArecurse) were performed with a range of infant-specific tuning parameters, which were selected in accordance with values tested in previous publications, to test parameter sensitivity. We tested iqr = [0.1, 0.5, 1.0, 1.5] for wavelet filtering and stdThresh = [10, 15, 20, 25] for tPCA. For tPCA, we used the following set of additional parameter thresholds: tMotion = 0.5 s, tMask = 1 s, ampThresh = 5, and nSV = 0.97 (97%) as evaluated by Yücel et al.[7] After implementation of each motion correction algorithm, we either directly calculated the data quality metrics or applied basic trial rejection to identify residual motion artifacts (using the motion artifact detection implemented in Homer2 as hmrMotionArtifact with parameter thresholds: tMotion = 1 s, tMask = 1 s, stdThresh = 25, and ampThresh = 1) and remove affected trials before calculating the HRF recovery error metric. In the fifth processing stream, we applied first tPCA and then wavelet filtering before metric calculation. We refer to consecutive implementation of both algorithms as "stacking." Stacking was performed testing a single combination of tuning parameters that performed best in processing stream three and four, namely stdThresh = 15 for tPCA and $iqr = 0.5$ for wavelet filtering.

For all processing streams, the resulting data were bandpass filtered, using a third-order Butterworth filter between 0.05 and 0.8 Hz to eliminate slow drift and cardiac artifact, followed by conversion of the OD data to change in chromophore concentration (using the modified Beer–Lambert law[18,19] with a DPF = 5[20]). Last, for calculation of the HRF recovery error metric, stimulus trials were averaged in a window from 2-s prestimulus onset to 16-s poststimulus onset using the prestimulus time for baseline correction to recover the mean HRF for every channel per dataset. For reasons of brevity, we only report oxyHb results although the deoxyHb results were similar (data not shown).

2.2.5 *Metrics for comparison*

Probe motion parameters. We quantified infant fNIRS probe/head motion for all age groups in the video conditions using an accelerometer. Accelerometer data were numerically integrated and combined over three dimensions to get a composite measure of head speed at each timepoint. Median and maximum head speed were calculated for the total time of video stimulus presentation for each subject, per age groups separately. A motion trace for an example 7-month-old participant is shown in Fig. 3.

Data quality metrics. We assessed data quality metrics of the calculated oxyHb channel timecourses for all age groups in the video and live conditions. We estimated and report median percent (percent noise) and range of percent of the dataset identified as motion (using the automated artifact detection hmrMotionArtifact with parameter thresholds: tMotion = 1 s, tMask = 1 s, stdThresh = 25, and ampThresh = 1). Moreover, we estimated: (i) within-subject standard deviation (stdev) and (ii) within-subject range (range) of the oxyHb channel timecourses for each dataset, per age groups separately in the video and live conditions, before and after motion corrections. We calculated the stdev and range of oxyHb

timecourses for each channel and reported median of within-subject stdev and range of oxyHb timecourses aggregated over channels. These data quality metrics should reflect motion artifacts but may also be impacted by other sources of noise in the data, such as physiological oscillations in cerebral hemodynamics.[21]

Hemodynamic response recovery error metric. To test accuracy of each motion correction method on the simulated data, we calculated the root-mean-squared-error (RMSE) of true, simulated HRF versus recovered HRF over the hemodynamic timecourse for each channel (and all chromophores), per age groups separately. We report the mean of the within-subject RMSE aggregated over channels.

Statistical analysis. Statistical testing was performed using SPSS version 24.0 (IBM SPSS Statistics for Mac, Armonk, New York: IBM Corp). For each metric, differences between age groups for video conditions were tested using a one-way analysis of variance (ANOVA) with between-subjects factor infant age (5 months versus 7 months versus 12 months) and probe motion parameters, uncorrected data quality metrics, data quality metrics after motion correction, and RMSE per motion correction method, respectively, as dependent variable. These analyses tested the hypotheses that age impacted motion during the experiment, data quality, or motion correction performance. Moreover, we computed Pearson's correlations between probe motion parameters and uncorrected data quality metrics and the RMSE per motion correction method over all age groups in the video conditions. These analyses explored whether the different coarse measures of participant movement were associated with each other and whether our measures of participant movement were associated with the HRF recovery error. If associations were found between probe motion parameters and HRF recovery error, it would indicate that the motion correction methods were unable to completely remove the impact of motion on the fNIRS signals. Differences between video and live conditions in age-similar samples (7 months versus 6 to 8 months) in the uncorrected data quality metrics were tested using two-tailed, two-sample *t*-tests. We expected infants to move more during the live interaction task than during video presentation and, therefore, to see a higher prevalence of motion artifacts and reduced data quality in the infant fNIRS timecourse data for live conditions. To quantitatively assess motion correction performance for all age groups in the video and live conditions, we compared uncorrected data quality metrics versus data quality metrics after motion correction using two-tailed, paired *t*-tests. We also tested accuracy of each motion correction method on the simulated data by comparing RMSE per motion correction method versus RMSE for no-motion correction (or trial rejection only) using two-tailed, paired *t*-tests. We expected a decrease in metric values, specifically e.g., smaller stdev and range of oxyHb timecourses or lower RMSE compared to no correction, to indicate better motion correction performance. However, it should be noted that small stdev and range may indicate that hemodynamic response features have been eliminated from the data along with noise. Therefore, RMSE was used to evaluate overall performance of motion correction algorithms, and the relationship between RMSE and the stdev and range was assessed to determine if these easily calculated metrics can be used as shorthand for a simulation study and also for comparison between video and live data.

3 Results

3.1 *Quantification of Infant fNIRS Probe Motion and Data Quality Metrics*

Infant probe motion during video presentation was quantified over the whole course of the experiment by finding the median and maximum head speed. Median and maximum head speed per age are shown in Fig. 4. Median head speed was not different between age groups [one-way ANOVA, $F(2,57) = 2.07$, $p > 0.14$], indicating no differences in overall probe/head motion during video presentation across infant ages. Age groups, however, did differ in maximum head speed [$F(2,57) = 5.28$, $p = 0.008$]. Five-month-olds showed significantly lower fastest head motion over the course of the video presentation than 7-month-olds [$t(38) = -2.96$, false-discovery-rate (FDR)-corr. $p \leq 0.01$] and 12-month-olds [$t(38) = -3.62$, FDR-corr. $p \leq 0.005$]. Seven- and 12-month-olds did not differ [$t(38) = 0.427$, FDR-corr. $p > 0.67$].

For 5-month-olds, overall median = 5.5% (range = 12.9%) of the dataset per subject was identified as motion; for 7-month-olds, overall median = 6.3% (range = 97.2%) of the dataset per subject was identified as motion; and for 12-month-olds, overall median = 8.9% (range = 99.5%) of the dataset per subject was identified as motion. Across age groups, no differences in percent noise [$F(2,57) = 2.053$, $p > 0.14$] were found. Infant fNIRS data quality was further quantified by estimating two coarse measures of variability in the timecourse data; specifically, we calculated within-subject standard deviation (stdev) and range of oxyHb timecourses. Data quality metrics

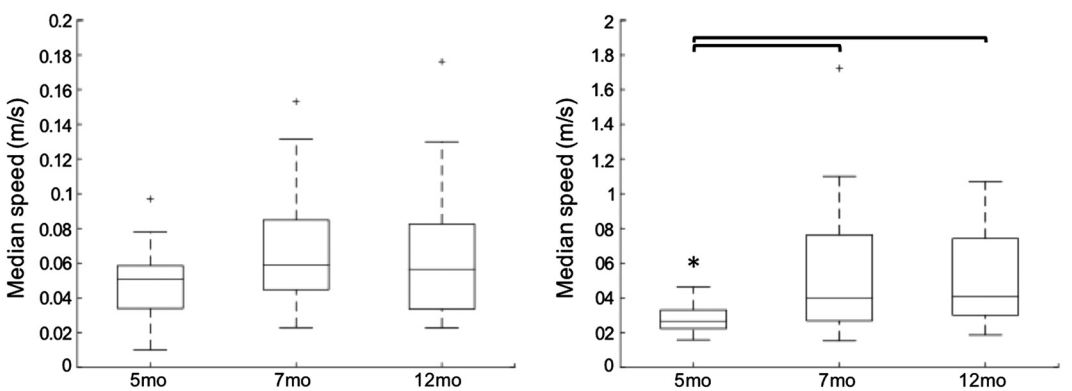

Fig. 4 Probe motion parameters for infant fNIRS data collected during video stimulus presentation across infant ages.

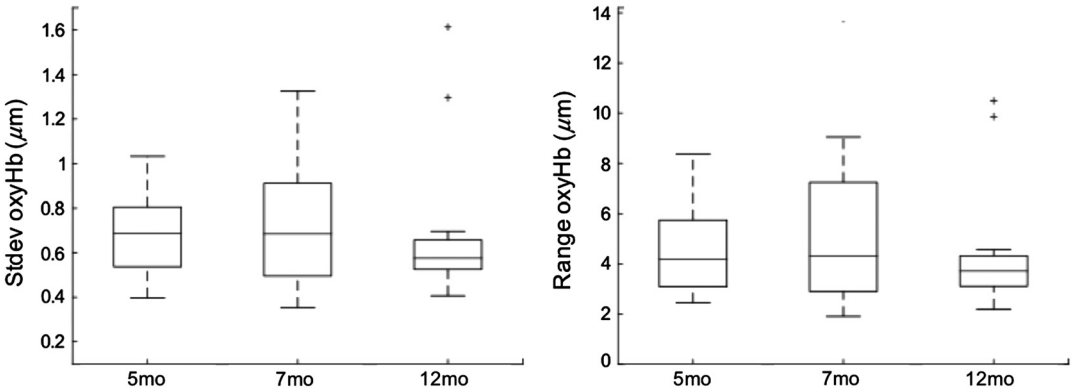

Fig. 5 Data quality metrics, i.e., stdev and range of oxyHb timecourses, for infant fNIRS data collected during video stimulus presentation across infant ages.

are shown in Fig. 5. No differences in either stdev [$F(2,57) = 0.335$, $p > 0.72$] or range [$F(2,57) = 0.968$, $p > 0.39$] of oxyHb timecourses were found, indicating no differences in uncorrected fNIRS data quality across infant ages.

3.1.1 *Link between infant fNIRS probe motion and data quality metrics*

We explored whether infant probe motion parameters and (uncorrected) fNIRS data quality metrics were linked. Pearson's correlations are presented in Table 1. Across age groups ($N = 60$) in video conditions, infant median and maximum head speed and stdev and range of oxyHb timecourses showed a moderate, positive correlation; however, correlations were nonsignificant after correction for multiple comparisons. Percent noise was not correlated with median and maximum head speed.

3.2 *Motion Correction Performance*

3.2.1 *Data quality metrics after motion correction*

Analyses of data quality metrics after motion correction (wavelet, tPCA, and stacking) were computed within age groups separately, results are presented in Table 2. Performance of motion correction algorithms and parameters was quantitatively assessed by comparing data quality metrics (stdev and range of oxyHb timecourses) before versus after motion correction. Data quality metrics per motion correction method are shown in Fig. 6. For all infant ages, wavelet filtering significantly reduced stdev and range oxyHb in all cases (all parameters, two-tailed, paired t-tests $ps < 0.001$). tPCA significantly reduced stdev and range oxyHb in all cases for 5- and 7-month-olds (all parameters, $ps \leq 0.001$ to 0.05). For 12-month-olds, tPCA with stdThresh 10 and 15 significantly reduced stdev and range oxyHb ($ps \leq 0.01$ to 0.03); however, tPCA with stdThresh 20 and 25 only by trend ($ps \leq 0.04$ to 0.08). Stacking, in all cases and for all age groups, significantly reduced stdev and range oxyHb (all parameters, $ps < 0.001$). Across infant ages, no differences in stdev oxyHb (one-way ANOVA, $ps > 0.20$) and range oxyHb (one-way ANOVA, $ps > 0.31$) per motion correction method were found.

3.2.2 *Simulated hemodynamic response recovery error metric*

Figure 7 shows the true, simulated HRF and the impact of the different motion correction methods on the recovered HRF for an example, 7-month-old participant and channel.

To assess accuracy of motion correction methods (no correction, trial rejection, wavelet, tPCA, and stacking), we compared true, simulated HRF versus recovered HRF and calculated the recovery error (RMSE) over the hemodynamic timecourses for each channel per dataset, within age groups separately. RMSE per motion correction method is shown in Fig. 8. Motion correction performance was quantitatively assessed by comparing RMSE per motion correction method versus no correction. Results are presented in Table 3. Trial rejection did not improve HRF recovery. In fact, for 5- and 12-month-olds, RMSE of basic trial rejection versus no correction did not differ (two-tailed paired t-test, $ps > 0.09$), and for 7-month-olds, the RMSE of basic trial rejection was significantly increased compared to no correction ($p = 0.015$). On the other hand, for all infant ages, wavelet filtering regardless of strictness of parameter yielded a significant decrease in RMSE compared to no correction in all cases ($ps \leq 0.001$ to 0.018). tPCA did not perform as consistently well as wavelet motion correction on the infant data. For 5- and 7-month-olds, stdThresh parameter 10 and 15 significantly decreased RMSE compared to no correction ($ps \leq 0.001$ to 0.01). For 12-month-olds, these parameters only by trend decreased the RMSE compared to no correction ($ps > 0.05$ to 0.07). In all cases, tPCA with stdThresh 20 and 25 did not improve HRF recovery ($ps > 0.13$ to 0.91). Stacking, however, yielded a significant decrease in RMSE compared to no correction in all cases and for all age groups ($ps \leq 0.001$ to 0.02). Across infant ages, no differences in HRF recovery error per motion correction method were found (one-way ANOVA, $ps > 0.18$).

3.2.3 *Link between infant fNIRS probe motion, data quality metrics, and hemodynamic response recovery error metric*

We explored whether infant fNIRS probe motion parameters and (uncorrected) data quality metrics were linked to the HRF recovery error. Pearson's correlations are presented in Table 1. Across age groups ($N = 60$) in video conditions, median and maximum head speed were not correlated with RMSE per motion correction method (FDR-corr. $ps > 0.05$). Percent noise as well as stdev and range of oxyHb uncorrected timecourse data, however, was significantly, positively correlated with RMSE per motion correction method (FDR-corr. $ps < 0.001$). This indicates that prevalence of motion artifacts and coarse measures of variability in the uncorrected fNIRS data, but less so coarse measures of probe/head motion, may function as

Table 1 Bivariate correlation. Linking probe motion parameters (median and maximum head speed) and data quality metrics (percent noise, within-subject standard deviation, and range of uncorrected oxyHb timecourse data) to hemodynamic response recovery error.

						Recovery metric (RMSE)										
	1	2	3	4	5	No correction	Trial rejection	Wavelet 0.1	Wavelet 0.5	Wavelet 1.0	Wavelet 1.5	tPCA 10	tPCA 15	tPCA 20	tPCA 25	Stacking
1 Median speed	—	—	—	—	—	$r = 0.200$, $p = 0.126$, $n = 60$	$r = 0.134$, $p = 0.329$, $n = 55$	$r = 0.218$, $p = 0.094$, $n = 60$	$r = 0.173$, $p = 0.191$, $n = 59$	$r = 0.184$, $p = 0.172$, $n = 57$	$r = 0.175$, $p = 0.193$, $n = 57$	$r = 0.212$, $p = 0.117$, $n = 56$	$r = 0.274$, $p = 0.041$, $n = 56$	$r = 0.230$, $p = 0.088$, $n = 56$	$r = 0.220$, $p = 0.103$, $n = 56$	$r = 0.173$, $p = 0.189$, $n = 59$
2 Maximum speed	$r = 0.771$, $p < 0.001$, $n = 60$	—	—	—	—	$r = 0.189$, $p = 0.148$, $n = 60$	$r = 0.160$, $p = 0.242$, $n = 55$	$r = 0.165$, $p = 0.209$, $n = 60$	$r = 0.093$, $p = 0.481$, $n = 59$	$r = 0.095$, $p = 0.483$, $n = 57$	$r = 0.093$, $p = 0.493$, $n = 57$	$r = 0.026$, $p = 0.848$, $n = 56$	$r = 0.111$, $p = 0.417$, $n = 56$	$r = 0.094$, $p = 0.491$, $n = 56$	$r = 0.193$, $p = 0.154$, $n = 56$	$r = 0.034$, $p = 0.797$, $n = 59$
3 Percent noise[a]	$r = 0.114$, $p = 0.401$, $n = 56$	$r = 0.190$, $p = 0.160$, $n = 56$	—	—	—	$r = 0.700$, $p < 0.001$, $n = 56$	$r = 0.750$, $p < 0.001$, $n = 55$	$r = 0.493$, $p < 0.001$, $n = 56$	$r = 0.564$, $p < 0.001$, $n = 56$	$r = 0.584$, $p \leq 0.001$, $n = 56$	$r = 0.600$, $p \leq 0.001$, $n = 56$	$r = 0.455$, $p \leq 0.001$, $n = 56$	$r = 0.589$, $p \leq 0.001$, $n = 56$	$r = 0.584$, $p \leq 0.001$, $n = 56$	$r = 0.674$, $p \leq 0.001$, $n = 56$	$r = 0.451$, $p \leq 0.001$, $n = 56$
4 Uncorrected standard deviation	$r = 0.260$, $p = 0.045$, $n = 60$	$r = 0.272$, $p = 0.036$, $n = 60$	$r = 0.643$, $p < 0.001$, $n = 56$	—	—	$r = 0.904$, $p < 0.001$, $n = 60$	$r = 0.815$, $p < 0.001$, $n = 55$	$r = 0.697$, $p < 0.001$, $n = 60$	$r = 0.731$, $p < 0.001$, $n = 59$	$r = 0.806$, $p < 0.001$, $n = 57$	$r = 0.777$, $p < 0.001$, $n = 57$	$r = 0.633$, $p < 0.001$, $n = 56$	$r = 0.604$, $p < 0.001$, $n = 56$	$r = 0.631$, $p < 0.001$, $n = 56$	$r = 0.762$, $p < 0.001$, $n = 56$	$r = 0.600$, $p < 0.001$, $n = 59$
5 Uncorrected range	$r = 0.226$, $p = 0.082$, $n = 60$	$r = 0.307$, $p = 0.017$, $n = 60$	$r = 0.630$, $p < 0.001$, $n = 56$	$r = 0.937$, $p < 0.001$, $n = 60$	—	$r = 0.876$, $p < 0.001$, $n = 60$	$r = 0.813$, $p < 0.001$, $n = 55$	$r = 0.680$, $p < 0.001$, $n = 60$	$r = 0.699$, $p < 0.001$, $n = 59$	$r = 0.738$, $p < 0.001$, $n = 57$	$r = 0.702$, $p < 0.001$, $n = 57$	$r = 0.580$, $p < 0.001$, $n = 56$	$r = 0.538$, $p < 0.001$, $n = 56$	$r = 0.580$, $p < 0.001$, $n = 56$	$r = 0.678$, $p < 0.001$, $n = 56$	$r = 0.526$, $p < 0.001$, $n = 59$

Notes: Results are reported uncorrected for multiple comparisons. RMSE = Root-mean-squared-error of true, simulated versus recovered hemodynamic responses, tPCA = targeted principal component analysis, stacking = consecutive application of first tPCA with stdTresh = 15, and then wavelet filtering with $iqr = 0.5$.
[a]Percent noise excluding four extreme outliers, identified as ± 3 standard deviations from group mean in univariate boxplots.

Table 2 Evaluation of data quality metrics before versus after motion correction for video conditions across infant ages.

	Data quality metrics	5 months		7 months		12 months		5 months versus 7 months versus 12 months
		Mean (stdev)	Paired *t*-test	Mean (stdev)	Paired *t*-test	Mean (stdev)	Paired *t*-test	*F*-test
Uncorrected data	Standard deviation	0.6990 (0.1863)	—	0.7238 (0.2721)	—	0.6592 (0.2859)	—	$F(2,57) = 0.335$, $p = 0.72$
	Range	4.5528 (1.7601)	—	5.2448 (2.9779)	—	4.2347 (2.1380)	—	$F(2,57) = 0.968$, $p = 0.39$
Wavelet 0.1	Standard deviation	0.3198 (0.0530)	$t(19) = 11.822$, $p < 0.001$	0.3217 (0.0848)	$t(19) = 9.238$, $p < 0.001$	0.3025 (0.0903)	$t(19) = 7.908$, $p < 0.001$	$F(2,57) = 0.368$, $p = 0.69$
	Range	1.5111 (0.2935)	$t(19) = 8.782$, $p < 0.001$	1.4917 (0.4306)	$t(19) = 6.405$, $p < 0.001$	1.4022 (0.4732)	$t(19) = 7.454$, $p < 0.001$	$F(2,57) = 0.408$, $p = 0.67$
Wavelet 0.5	Standard deviation	0.4135 (0.0759)	$t(19) = 10.218$, $p < 0.001$	0.4062 (0.1086)	$t(19) = 7.948$, $p < 0.001$	0.3816 (0.1016)	$t(19) = 6.526$, $p < 0.001$	$F(2,57) = 0.603$, $p = 0.55$
	Range	1.9679 (0.3970)	$t(19) = 7.674$, $p < 0.001$	1.9427 (0.5869)	$t(19) = 5.915$, $p < 0.001$	1.8421 (0.5815)	$t(19) = 6.677$, $p < 0.001$	$F(2,57) = 0.316$, $p = 0.73$
Wavelet 1.0	Standard deviation	0.5110 (0.0984)	$t(19) = 8.122$, $p < 0.001$	0.4900 (0.1382)	$t(19) = 6.636$, $p < 0.001$	0.4643 (0.1195)	$t(19) = 5.037$, $p < 0.001$	$F(2,57) = 0.762$, $p = 0.47$
	Range	2.5269 (0.5699)	$t(19) = 6.499$, $p < 0.001$	2.4173 (0.7213)	$t(19) = 5.232$, $p < 0.001$	2.3007 (0.6225)	$t(19) = 5.446$, $p < 0.001$	$F(2,57) = 0.623$, $p = 0.54$
Wavelet 1.5	Standard deviation	0.5730 (0.1187)	$t(19) = 6.559$, $p < 0.001$	0.5463 (0.1579)	$t(19) = 5.537$, $p < 0.001$	0.5181 (0.1464)	$t(19) = 4.336$, $p < 0.001$	$F(2,57) = 0.747$, $p = 0.48$
	Range	2.9056 (0.6593)	$t(19) = 5.549$, $p < 0.001$	2.8034 (0.8972)	$t(19) = 4.699$, $p < 0.001$	2.6100 (0.7491)	$t(19) = 4.919$, $p < 0.001$	$F(2,57) = 0.751$, $p = 0.48$
tPCA 10	Standard deviation	0.4455 (0.0879)	$t(19) = 5.471$, $p < 0.001$	0.4868 (0.1339)	$t(19) = 4.129$, $p = 0.001$	0.5037 (0.0966)	$t(19) = 2.792$, $p = 0.012$	$F(2,57) = 1.534$, $p = 0.22$
	Range	2.8158 (0.9492)	$t(19) = 3.800$, $p = 0.001$	3.4464 (1.6529)	$t(19) = 2.660$, $p = 0.015$	3.2699 (1.2826)	$t(19) = 2.373$, $p = 0.028$	$F(2,57) = 1.203$, $p = 0.31$
tPCA 15	Standard deviation	0.5669 (0.1084)	$t(19) = 4.300$, $p < 0.001$	0.5406 (0.1216)	$t(19) = 4.020$, $p = 0.001$	0.5469 (0.1136)	$t(19) = 2.426$, $p = 0.025$	$F(2,57) = 0.286$, $p = 0.75$
	Range	3.2959 (0.7156)	$t(19) = 3.702$, $p = 0.002$	3.2355 (1.0054)	$t(19) = 3.627$, $p = 0.002$	3.4040 (1.0054)	$t(19) = 2.537$, $p = 0.020$	$F(2,57) = 0.161$, $p = 0.85$
tPCA 20	Standard deviation	0.6344 (0.1434)	$t(19) = 4.513$, $p < 0.001$	0.6138 (0.1744)	$t(19) = 3.641$, $p = 0.002$	0.5723 (0.1176)	$t(19) = 1.985$, $p = 0.062$	$F(2,57) = 0.924$, $p = 0.40$
	Range	3.7222 (0.9751)	$t(19) = 3.227$, $p = 0.004$	3.8528 (1.4595)	$t(19) = 3.217$, $p = 0.005$	3.5224 (1.0617)	$t(19) = 2.275$, $p = 0.035$	$F(2,57) = 0.395$, $p = 0.68$
tPCA 25	Standard deviation	0.0.6704 (0.1602)	$t(19) = 2.357$, $p = 0.029$	0.6465 (0.2002)	$t(19) = 3.287$, $p = 0.004$	0.6097 (0.1820)	$t(19) = 1.864$, $p = 0.078$	$F(2,57) = 0.568$, $p = 0.57$
	Range	4.0432 (1.1222)	$t(19) = 2.101$, $p = 0.049$	4.1893 (1.8248)	$t(19) = 2.953$, $p = 0.008$	3.7792 (1.3781)	$t(19) = 1.890$, $p = 0.074$	$F(2,57) = 0.400$, $p = 0.67$
Stacking	Standard deviation	0.3667 (0.0612)	$t(19) = 9.249$, $p < 0.001$	0.3401 (0.0512)	$t(19) = 7.151$, $p < 0.001$	0.3381 (0.0544)	$t(19) = 5.889$, $p < 0.001$	$F(2,57) = 1.640$, $p = 0.20$
	Range	1.7233 (0.2979)	$t(19) = 7.643$, $p < 0.001$	1.6110 (0.2947)	$t(19) = 5.813$, $p < 0.001$	1.6153 (0.2727)	$t(19) = 6.111$, $p < 0.001$	$F(2,57) = 0.973$, $p = 0.38$

Notes: Results are reported uncorrected for multiple comparisons. Stdev = Standard deviation, tPCA = targeted principal component analysis, stacking = consecutive application of first tPCA with stdTresh = 15 and then wavelet filtering with $iqr = 0.5$.

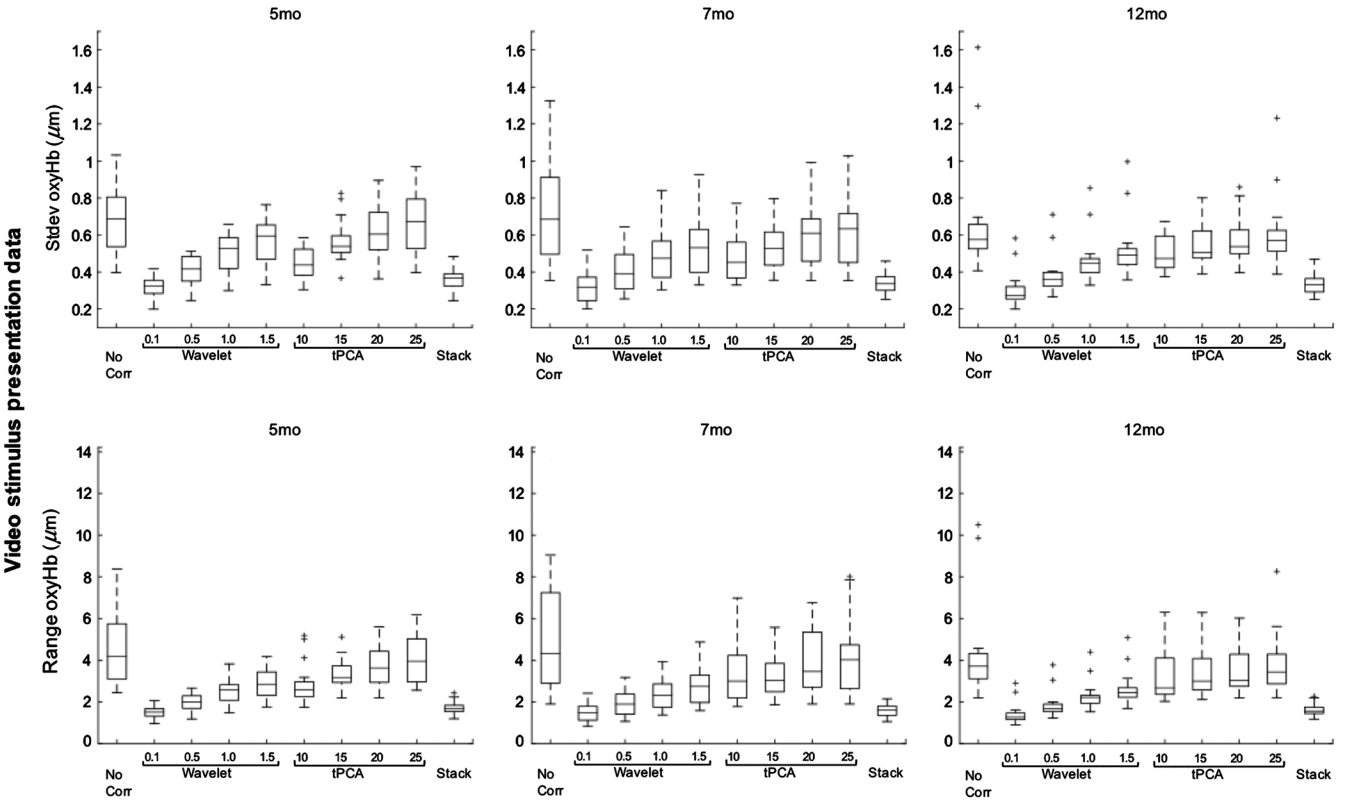

Fig. 6 Data quality metrics before versus after motion correction (wavelet, tPCA, and stacking) for video timecourse data across infant ages.

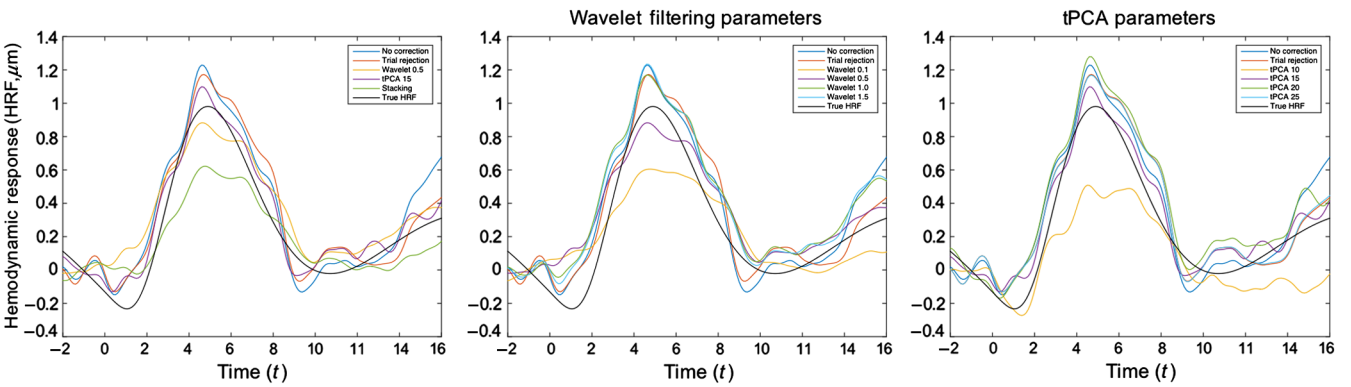

Fig. 7 True, simulated HRF (in black) and the impact of different motion correction methods (no correction, trial rejection, wavelet, tPCA, and stacking) on the recovered HRF for an example 7-month-old participant and channel.

a proxy for accuracy of HRF recovery when using motion correction.

3.3 *Infant fNIRS Data Collected During Live Stimulus Presentation*

3.3.1 *Data quality metrics*

Differences in infant fNIRS timecourse data collected during live versus video stimulus presentation were examined by estimating and comparing data quality metrics (specifically, e.g., percent noise, within-subject stdev, and range of the oxyHb timecourses) in age-similar samples. Data quality metrics for live interaction versus video presentation are shown in Fig. 9. For infant data collected during live stimulus presentation, overall median = 4.1% (range = 11.2%) of the dataset per subject was identified as motion, which was not different from percent noise identified in 7-month-old infant data collected during video presentation [two-tailed, two-sample t-test $t(28) = 1.046$, $p > 0.30$]. Further, in age-similar samples, no differences in stdev [$t(28) = -0.346$, $p > 0.73$] or range [$t(28) = 0.288$, $p > 0.77$] oxyHb between infant live interaction and video data were found. This indicates that prevalence of motion artifacts and data quality did not differ dependent on method of stimulus presentation.

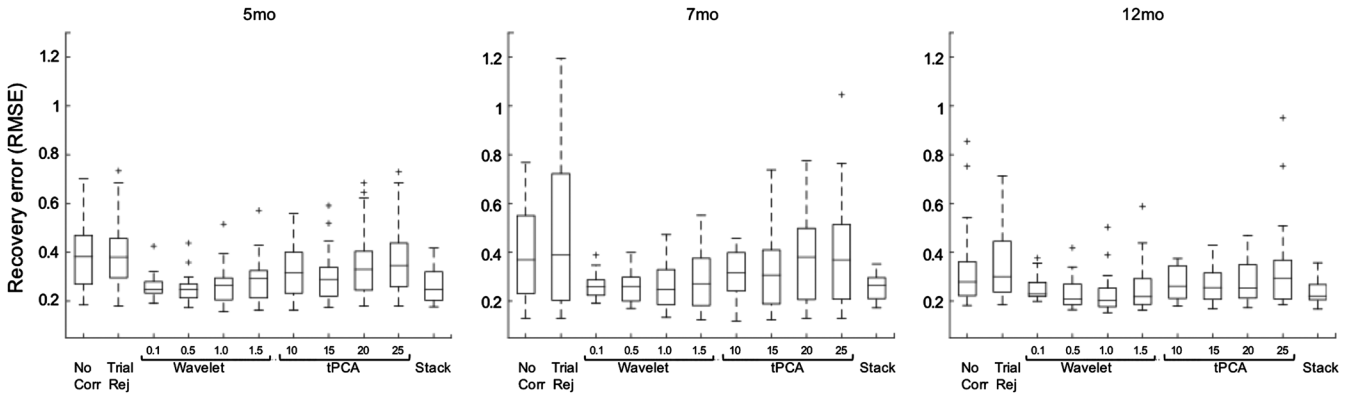

Fig. 8 Hemodynamic response recovery error metric per motion correction method (no correction, trial rejection, wavelet, tPCA, and stacking) for video timecourse data combined with simulated hemodynamic responses across infant ages.

3.3.2 *Data quality metrics after motion correction*

Performance of motion correction algorithms and parameters (wavelet, tPCA, and stacking) for live interaction data was quantitatively assessed by comparing data quality metrics before versus after motion correction, results are presented in Table 4; data quality metrics per motion correction algorithm and parameter are shown in Fig. 10. For live interaction data, wavelet filtering significantly reduced stdev and range oxyHb in all cases (all parameters, two-tailed, paired t-tests $ps < 0.001$). tPCA with stdThresh 10, 15, and 20 also significantly reduced stdev and range oxyHb ($ps \leq 0.001$ to 0.04); however, tPCA with stdThresh 25 did not ($ps > 0.11$ to 0.14). Stacking significantly reduced stdev and range oxyHb in all cases ($ps < 0.001$). Taken together, this shows that motion correction in almost all cases significantly reduced stdev and range in infant fNIRS data collected during live stimulus presentation, indicating motion correction was well suited and performed similarly well as for infant fNIRS data collected during video stimulus presentation.

4 Discussion

4.1 *Infant fNIRS Probe Motion and Data Quality*

Infants' head motion during a standard infant fNIRS task, i.e., 2-min video presentation, over the first year of life was shown to be fairly constant as reflected in the median speed metric but was slow. Median head motion did not change with infant age. However, 7- and 12-month-old infants showed higher maximum head speed over the course of the video presentation compared to 5-month-old infants, which could be due to increasing motor development in the second half of the first year of life. Notably, percent noise as well as the within-subject standard deviation and range of the fNIRS timecourse data did not change with infant age, indicating that age-dependent differences in fastest head motion during video presentation did not impact uncorrected fNIRS data quality. Specifically, we did not see more motion artifacts, such as spikes, in the uncorrected timecourse data for older infant participants. While we found measured probe motion parameters and uncorrected data quality metrics to be positively correlated, our findings mitigate concerns that differences in maximum speed in older compared to younger infant ages might result in more motion artifacts and poor data quality in older infants.

4.2 *Motion Correction Performance*

We showed that on average over different groups of infant ages using any motion correction algorithm was better than using no correction or basic trial rejection at eliminating motion artifacts and improving hemodynamic response recovery. This finding was not dependent on infant age, indicating that motor development over the first year of life did not affect motion correction performance. Moreover, we found that metrics calculated from the infant fNIRS timecourse data themselves were most informative of hemodynamic response recovery performance. While we found a positive correlation between these data quality metrics and measured motion parameters in the uncorrected timecourse data, median or maximum head speed did not relate to hemodynamic response recovery performance after motion correction. This finding suggests that it could be sufficient to correct fNIRS data using motion correction algorithms that do not include additional motion parameters instead of using methods derived from simultaneous collection of accelerometer data (such as Refs. 22 and 23), reducing experimental complications for infant studies.

Trial rejection alone did not improve hemodynamic response recovery as compared to using no correction at all. This finding is particularly important since basic trial rejection is one of the most commonly used motion correction approaches for fNIRS data.[6] Yet, it does not seem well suited for infant data. Previous simulation studies testing motion correction methods for adult data[4,6] also show that trial rejection does not perform well and may fail to improve hemodynamic response recovery. While basic trial rejection is still widely used, especially in infant fNIRS research, it performed no better than not applying motion correction at all. Hence, we do not recommend using basic trial rejection as the only motion correction strategy independent of sample characteristic.

On the other hand, any motion correction algorithm should perform roughly the same for groups of different infant ages and will in almost all cases significantly improve hemodynamic response recovery compared to using no-motion correction at all. Our recommendations do not appear to differ from recommendations for other, older developmental[9] or adult populations.[4,6,7]

Wavelet filtering, regardless of infant age, seems to perform well. We found overall good performance of wavelet motion correction for infant data and suggest using wavelet filtering

Table 3 Evaluation of hemodynamic response recovery error per motion correction method versus no correction across infant ages.

	Recovery metric	5 months		7 months		12 months		5 months versus 7 months versus 12 months
		Mean (stdev)	Paired *t*-test	Mean (stdev)	Paired *t*-test	Mean (stdev)	Paired *t*-test	*F*-test
No correction	RMSE	0.3803 (0.1323)	—	0.4045 (0.2034)	—	0.3417 (0.1836)	—	$F(2, 57) = 0.651$, $p = 0.53$
Trial rejection	RMSE	0.4084 (0.1651)	$t(19) = -1.279$, $p = 0.216$	0.4916 (0.3341)	$t(17) = -2.720$, $p = 0.015$	0.4062 (0.2868)	$t(16) = -1.781$, $p = 0.094$	$F(2,52) = 0.599$, $p = 0.55$
Wavelet 0.1	RMSE	0.2593 (0.0505)	$t(19) = 5.278$, $p < 0.001$	0.2652 (0.0518)	$t(19) = 3.887$, $p = 0.001$	0.2506 (0.0496)	$t(19) = 2.686$, $p = 0.010$	$F(2,57) = 0.417$, $p = 0.66$
Wavelet 0.5	RMSE	0.2534 (0.0617)	$t(19) = 6.352$, $p < 0.001$	0.2583 (0.0680)	$t(18) = 4.622$, $p < 0.001$	0.2340 (0.0679)	$t(19) = 3.946$, $p = 0.001$	$F(2,56) = 0.749$, $p = 0.48$
Wavelet 1.0	RMSE	0.2713 (0.0851)	$t(19) = 6.804$, $p < 0.001$	0.2706 (0.0971)	$t(18) = 5.159$, $p < 0.001$	0.2315 (0.0903)	$t(17) = 4.219$, $p = 0.001$	$F(2,54) = 1.161$, $p = 0.32$
Wavelet 1.5	RMSE	0.2959 (0.0985)	$t(19) = 5.996$, $p < 0.001$	0.2932 (0.1205)	$t(18) = 5.166$, $p < 0.001$	0.2646 (0.1144)	$t(17) = 2.606$, $p = 0.018$	$F(2,54) = 0.452$, $p = 0.64$
tPCA 10	RMSE	0.3264 (0.1029)	$t(19) = 2.858$, $p = 0.010$	0.3084 (0.1023)	$t(18) = 3.441$, $p = 0.003$	0.2689 (0.0686)	$t(16) = 1.924$, $p = 0.072$	$F(2,53) = 1.780$, $p = 0.18$
tPCA 15	RMSE	0.3071 (0.1101)	$t(19) = 4.692$, $p < 0.001$	0.3254 (0.1545)	$t(18) = 2.736$, $p = 0.014$	0.2734 (0.0880)	$t(16) = 2.078$, $p = 0.054$	$F(2,53) = 0.838$, $p = 0.44$
tPCA 20	RMSE	0.3502 (0.1482)	$t(19) = 1.381$, $p = 0.183$	0.3729 (0.1881)	$t(18) = 1.583$, $p = 0.131$	0.2899 (0.1008)	$t(16) = 1.508$, $p = 0.151$	$F(2,53) = 1.422$, $p = 0.25$
tPCA 25	RMSE	0.3742 (0.1498)	$t(19) = 0.323$, $p = 0.750$	0.4086 (0.2394)	$t(18) = 0.119$, $p = 0.906$	0.3544 (0.2094)	$t(16) = -0.805$, $p = 0.433$	$F(2,53) = 0.336$, $p = 0.72$
Stacking	RMSE	0.2667 (0.0709)	$t(19) = 6.039$, $p < 0.001$	0.2584 (0.0568)	$t(19) = 4.005$, $p = 0.001$	0.2344 (0.0487)	$t(19) = 3.344$, $p = 0.003$	$F(2,56) = 1.535$, $p = 0.22$

Notes: Results are reported uncorrected for multiple comparisons. RMSE = Root-mean-squared-error of true, simulated versus recovered hemodynamic responses, Stdev = standard deviation, tPCA = targeted principal component analysis, stacking = consecutive application of first tPCA with stdTresh = 15 and then wavelet filtering with $iqr = 0.5$.

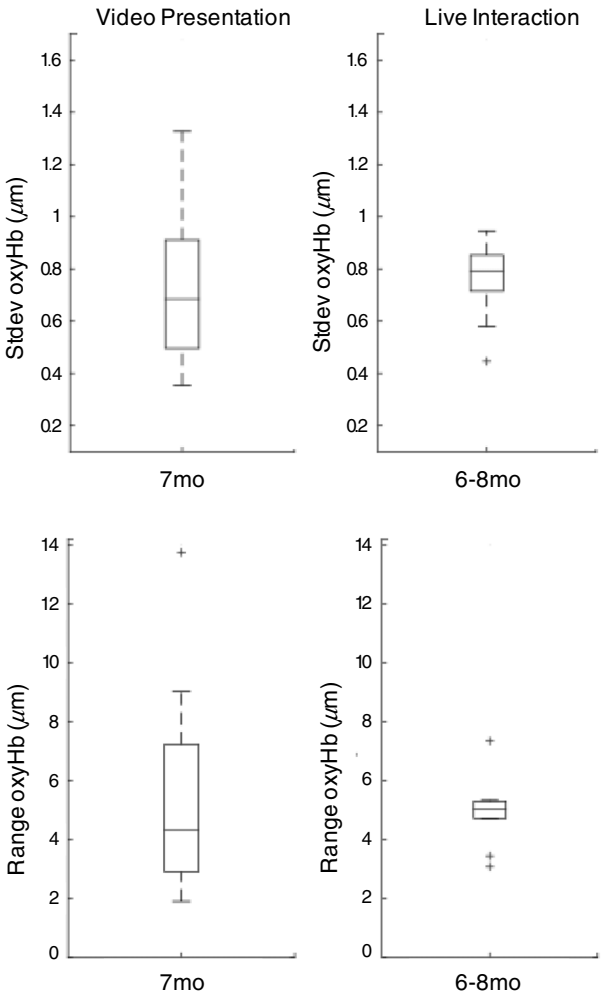

Fig. 9 Data quality metrics, i.e., stdev and range of oxyHb timecourses, for infant fNIRS data collected during live versus video stimulus presentation comparing age-similar samples.

with an *iqr* of 0.5 as a balance between eliminating noise in infant data and underestimating the magnitude of the recovered hemodynamic response; however, a range of parameter values seemed acceptable. Further, using wavelet motion correction may perform better than using tPCA, especially for 12-month-old infants.

Performance of tPCA motion correction for infant data overall was more sensitive to choice of parameter value than wavelet motion correction. We recommend using tPCA with a stdThresh of 15, but stdThresh parameter values between 10 and 15 seemed generally acceptable. While tPCA yielded slightly more nuanced effects, in most cases, it did perform well for infant data, except maybe for 12-month-old infant data. It remains, however, unclear as to why tPCA was not as well suited for 12-month-old infant data. Differences in elements of participant movement, such as maximum head speed, that could potentially have caused tPCA to perform worse, were not significantly related to the hemodynamic response recovery error, and 12-month-olds did not differ from other infant age groups in prevalence of spike motion artifacts and uncorrected within-subject standard deviation or range in the oxyHb timecourses, which relate to the hemodynamic response recovery error. With regard to tPCA in particular, the current findings raise the question as to whether tPCA might perform even better, if the standard deviation of timecourse data itself was used as parameter threshold, instead of using a parameter that is based on a multiplicative factor of the standard deviation. This approach would place a threshold on acceptable standard deviations in the OD data instead of allowing timecourses with uniformly high standard deviations throughout.

We found no specific harm or benefit to stacking motion correction algorithms for infant data, e.g., as tested in this work by applying first tPCA and then wavelet filtering. Stacking removes data identified as motion and makes the timecourse data look cleaner by lowering the standard deviation and range, nonetheless low standard deviation and range of oxyHb timecourses may indicate that hemodynamic response features have been eliminated from the data along with noise and the hemodynamic response recovery when stacking was not improved compared to wavelet filtering (or tPCA) only.

Taken together, the general risk for underestimation of the hemodynamic response is higher using stricter tuning parameters and this is even more so the case when stacking, which demonstrates the more general challenge of finding a good balance between removing the most noise in the data versus preserving the most signal when choosing motion correction strategies and parameters. We found that features of the uncorrected timecourse data, such as the within-subject standard deviation of oxyHb timecourses, relate significantly to the hemodynamic response recovery error and, thus, could be used as proxy for how well motion correction may perform on the data. We suggest that uncorrected standard deviation of the timecourse data could be used to guide the decision of how strict the motion correction and parameter values should be.

4.3 *Motion Correction for Live Interaction Data*

Overall, infant data collected during live stimulus presentation may not need to be treated differently when correcting for participant movement than infant data collected during more standard experimental settings, such as during video stimulus presentation. While we did not directly compare probe/head motion parameters between live and video conditions, we found no differences in data quality metrics of oxyHb timecourse data indicating method of stimulus presentation did not differentially impact prevalence of motion artifacts in the data or data quality in general. This finding is particularly interesting as we also analyzed participant physiology during live and video stimulus presentation in age-similar samples [specifically, we extracted heart rate (HR) responses from the fNIRS recordings using an established algorithm for infant data[24]], and we found significant differences in absolute HR between method of stimulus presentation but no differences in HR responses to stimuli. Infant absolute HR during live face-to-face interaction with the mother was significantly higher than during video presentation, which could be indicating higher overall physiological arousal during the live design.[25–27] Interestingly, variability in infant overall physiological arousal over the course of the experiment was fairly consistent and this did not vary dependent on method of stimulus presentation. While we see higher levels of infant overall physiological arousal during live stimulus presentation, which could cause more participant movement, we did not find higher levels of percent noise or greater within-subject standard deviation and range in the uncorrected live fNIRS timecourse data. While there is no accelerometer data from the live design, we speculate that the infants may have

Table 4 Evaluation of data quality metrics before versus after motion correction for live interaction data.

	Data quality metrics	Mean (stdev)	Paired *t*-test
Uncorrected data	Standard deviation	0.7558 (0.1474)	—
	Range	4.9512 (1.1549)	—
Wavelet 0.1	Standard deviation	0.3161 (0.0684)	$t(9) = 13.646$, $p < 0.001$
	Range	1.5646 (0.3780)	$t(9) = 10.022$, $p < 0.001$
Wavelet 0.5	Standard deviation	0.4241 (0.0788)	$t(9) = 12.478$, $p < 0.001$
	Range	2.0880 (0.4589)	$t(9) = 8.823$, $p < 0.001$
Wavelet 1.0	Standard deviation	0.5396 (0.0968)	$t(9) = 10.803$, $p < 0.001$
	Range	2.6889 (0.5406)	$t(9) = 7.745$, $p < 0.001$
Wavelet 1.5	Standard deviation	0.6205 (0.1121)	$t(9) = 7.640$, $p < 0.001$
	Range	3.1733 (0.6417)	$t(9) = 5.899$, $p < 0.001$
tPCA 10	Standard deviation	0.4433 (0.1807)	$t(9) = 7.667$, $p < 0.001$
	Range	3.2700 (1.6997)	$t(9) = 2.863$, $p = 0.019$
tPCA 15	Standard deviation	0.6055 (0.1483)	$t(9) = 5.741$, $p < 0.001$
	Range	3.6253 (0.9045)	$t(9) = 4.443$, $p = 0.002$
tPCA 20	Standard deviation	0.7060 (0.1298)	$t(9) = 2.688$, $p = 0.025$
	Range	4.2058 (0.7530)	$t(9) = 2.413$, $p = 0.039$
tPCA 25	Standard deviation	0.7199 (0.1397)	$t(9) = 1.799$, $p = 0.106$
	Range	4.4158 (0.8300)	$t(9) = 1.623$, $p = 0.139$
Stacking	Standard deviation	0.3495 (0.0920)	$t(9) = 12.991$, $p < 0.001$
	Range	1.7635 (0.4839)	$t(9) = 9.586$, $p < 0.001$

Notes: Results are reported uncorrected for multiple comparisons. Stdev = Standard deviation, tPCA = targeted principal component analysis, stacking = consecutive application of first tPCA with stdTresh = 15 and then wavelet filtering with $iqr = 0.5$.

found the live interaction more engaging and remained still for longer periods during the recording leading to high-quality live data. The current findings highlight the need for thoughtful experimental designs in infant studies, especially with regard to good headgear design [e.g., optimal optode-scalp coupling (see also Ref. 7)] to be able to collect usable data, and underscore in particular that live stimulus presentation designs in infants can be done without increasing noise in the data as compared to standard experimental designs using video stimulus presentation.

Moreover, this work is the first to support that conventional motion correction methods are well suited for infant fNIRS data collected during live stimulus presentation. Overall, using motion correction algorithms significantly lowered within-subject standard deviation and range in the oxyHb timecourse data after motion correction. We found that smaller uncorrected standard deviation in the oxyHb timecourse data relates significantly to better hemodynamic response recovery performance suggesting that using motion correction algorithms for live interaction data would also improve hemodynamic response recovery.

4.4 *Limitations and Future Directions*

We have chosen not to adjust motion correction parameters on an individual participant or channel level but to focus on hemodynamic response recovery for group analysis. On a group level, it is better to preserve as much data (and trials) as possible and to make trial-level adjustments. However, if the goal of the fNIRS study is not to report group results but instead to examine individual differences, then motion correction algorithm evaluation may need to be approached differently. This should be addressed in future simulation studies. Further, with this work, we cannot answer the question of whether motion correction would impact the statistical inferences made from the data. Future work could address this by testing trial rejection of random trials versus noisy trials (in addition to comparing versus no correction) to better understand how rejection of trials might impact analysis power. For example, Emberson et al.[28] found that while short channel information (from surface vasculature in infants) significantly impacted information in long channels, removing it did not result in significant differences in statistical inferences made from the experiment. Yet, removing short channel

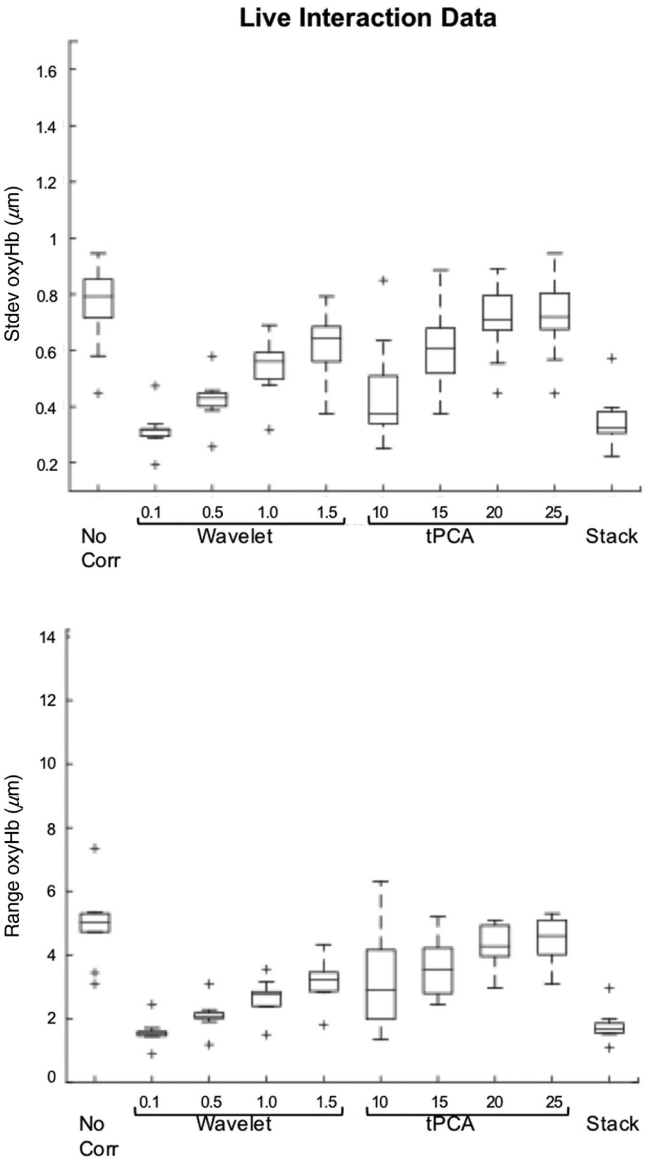

Fig. 10 Data quality metrics before versus after motion correction (wavelet, tPCA, and stacking) for infant fNIRS data collected during live stimulus presentation.

information may be more important when employing other experiments, e.g., task designs, including live stimulus presentation, during which infant overall physiological arousal can be significantly higher, and hence surface vasculature may significantly confound the functional brain signal.

While we have attempted to conduct a simulation analysis with conditions that are typically used in infant data, other approaches to evaluating motion correction for infant data could yield complementary useful information. Future directions could include looking at the shape of the HRF itself for example as implemented by Hu et al.[9] and Brigadoi et al.;[4] however, this approach could be complicated in live interaction data by potential differences in the shape of the HRF due to age/development and the additional physiological arousal associated with live stimulus presentation. Other potential useful analyses would include testing motion correction methods using the accelerometer timecourse instead of the summary statistic for hemodynamic response recovery, as done by Virtanen et al.,[22] or testing the performance of different stacking strategies and sensitivity of parameters.

To the best of our knowledge, this work is the first simulation study testing the performance of conventional motion correction methods for infant data. In sum, present results yield that motion correction algorithms, such as wavelet filtering and tPCA, perform well using infant-specific parameters and parameters may be used without fine-tuning for infant age or method of stimulus presentation. Live stimulus presentation constitutes a development in the field and a promising application of fNIRS, which promotes investigation of the early developing brain in a more naturalistic context, such as the early mother-infant interaction, an advantage no other infant-friendly imaging technique can offer.

Disclosures

The authors declare that they have no competing financial interest and no other potential conflicts of interest to disclose.

Acknowledgments

We thank the infants and their families for their participation. This work was supported by the National Institute of Mental Health Grant No. MH078829 awarded to C.A.N. Assistance with data collection was provided by Alissa Westerlund, Julia Cataldo, Anna Zhou, and Anna Fasman. The video stimulus was developed at the UW Autism Center of Excellence (No. P50HD055782) by Dr. Sarah Webb and colleagues. The live interaction task was developed at the University Hospital RWTH Aachen, Germany, by H.F.B., C.F., and Dr. Kerstin Konrad. The work was supported by a START grant awarded to C.F. and by a Fulbright grant awarded to H.F.B. The purchase of the Hitachi NIRS system for the University Hospital Aachen was supported by funding from the German Research Foundation DFG (INST 948/18-1 FUGG) awarded to K.K. Assistance with data collection was provided by Lea Jahnen and Wolfgang Scharke.

References

1. R. E. Vanderwert and C. A. Nelson, "The use of near-infrared spectroscopy in the study of typical and atypical development," *NeuroImage* **85**, 264–271 (2014).
2. 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).
3. R. N. Aslin, M. Shukla, and L. L. Emberson, "Hemodynamic correlates of cognition in human infants," *Annu. Rev. Psychol.* **66**, 349–379 (2015).
4. S. Brigadoi et al., "Motion artifacts in functional near-infrared spectroscopy: a comparison of motion correction techniques applied to real cognitive data," *NeuroImage* **85**(1), 181–191 (2014).
5. A. M. Chiarelli et al., "A kurtosis-based wavelet algorithm for motion artifact correction of fNIRS data," *NeuroImage* **112**, 128–137 (2015).
6. R. J. Cooper et al., "A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy," *Front. Neurosci.* **6**, 147 (2012).
7. M. A. Yücel et al., "Targeted principle component analysis: a new motion artifact correction approach for near-infrared spectroscopy," *J. Innov. Opt. Health Sci.* **7**(2), 1350066 (2014).
8. B. Molavi and G. A. Dumont, "Wavelet-based motion artifact removal for functional near-infrared spectroscopy," *Physiol. Meas.* **33**(2), 259–270 (2012).
9. X. S. Hu et al., "Comparison of motion correction techniques applied to functional near-infrared spectroscopy data from children," *J. Biomed. Opt.* **20**(12), 126003 (2015).

10. S. Lloyd-Fox et al., "Are you talking to me? Neural activations in 6-month-old infants in response to being addressed during natural interactions," *Cortex* **70**, 35–48 (2015).
11. J. Egetemeir et al., "Exploring the neural basis of real-life joint action: measuring brain activation during joint table setting with functional near-infrared spectroscopy," *Front. Hum. Neurosci.* **5**, 95 (2011).
12. J. Hirsch et al., "Frontal temporal and parietal systems synchronize within and across brains during live eye-to-eye contact," *NeuroImage* **157**, 314–330 (2017).
13. M. Suda et al., "Frontopolar activation during face-to-face conversation: an in situ study using near-infrared spectroscopy," *Neuropsychologia* **48**(2), 441–447 (2010).
14. M. M. Ravicz et al., "Infants' neural responses to facial emotion in the prefrontal cortex are correlated with temperament: a functional near-infrared spectroscopy study," *Front. Psychol.* **6**, 922 (2015).
15. H. F. Behrendt et al., "Postnatal mother-to-infant attachment in subclinically depressed mothers: dyads at risk?" *Psychopathology* **49**(4), 269–276 (2016).
16. Y. Zhang et al., "Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging," *J. Biomed. Opt.* **10**(1), 011014 (2005).
17. T. J. Huppert et al., "HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain," *Appl. Opt.* **48**(10), D280–D298 (2009).
18. M. Cope and D. T. Delpy, "System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination," *Med. Biol. Eng. Comput.* **26**(3), 289–294 (1988).
19. D. T. Delpy et al., "Estimation of optical pathlength through tissue from direct time of flight measurement," *Phys. Med. Biol.* **33**(12), 1433–1442 (1988).
20. 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).
21. M. A. Yücel et al., "Mayer waves reduce the accuracy of estimated hemodynamic response functions in functional near-infrared spectroscopy," *Biomed. Opt. Express* **7**(8), 3078–3088 (2016).
22. J. Virtanen et al., "Accelerometer-based method for correcting signal baseline changes caused by motion artifacts in medical near-infrared spectroscopy," *J. Biomed. Opt.* **16**(8), 087005 (2011).
23. A. Blasi et al., "Automatic detection of motion artifacts in infant functional optical topography studies," *Adv. Exp. Med. Biol.* **662**, 279–284 (2010).
24. K. L. Perdue et al., "Extraction of heart rate from functional near-infrared spectroscopy in infants," *J. Biomed. Opt.* **19**(6), 067010 (2014).
25. M. P. Fracasso, S. W. Porges, and M. E. Lamb, "Cardiac activity in infancy: reliability and stability of individual differences," *Infant Behav. Dev.* **17**(3), 277–284 (1994).
26. K. L. Perdue et al., "Differing developmental trajectories in heart rate responses to speech stimuli in infants at high and low risk for autism spectrum disorder," *J. Autism Dev. Disord.* **47**(8), 2434–2442 (2017).
27. J. E. Richards and B. J. Casey, "Heart rate variability during attention phases in young infants," *Psychophysiology* **28**(1), 43–53 (1991).
28. L. L. Emberson et al., "Isolating the effects of surface vasculature in infant neuroimaging using short-distance optical channels: a combination of local and global effects," *Neurophotonics* **3**(3), 031406 (2016).

Hannah F. Behrendt received her diploma in psychology from the Johannes Gutenberg-Universität Mainz, Germany, in September 2013. She is currently completing her PhD in clinical neuroscience in the Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy at the University Hospital RWTH Aachen, Germany, while also working as a postdoctoral research fellow at Boston Children's Hospital. Her research interests focus on neurobiological mechanisms underlying early mother–child interaction and developing imaging methods for application of live designs in pediatric populations.

Christine Firk is currently working as a postdoctoral research fellow in the Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy at the University Hospital RWTH Aachen. Her research interests focus on neurobiological mechanisms underlying early mother–child interaction. She is a licensed child psychologist. She received her master's degree in developmental psychology from Maastricht University in 2005 and received her PhD in psychology from Maastricht University in 2009.

Charles A. Nelson III is a professor of pediatrics and neuroscience and a professor of psychology in psychiatry at Harvard Medical School and holds the Richard David Scott Chair in Pediatric Developmental Medicine Research, Boston Children's Hospital. The Nelson Laboratory conducts research on a variety of problems in developmental cognitive neuroscience, with a particular focus on the effects of early experience on brain development and the development of autism. He has published over 300 peer-reviewed papers.

Katherine L. Perdue received her BS degree in physics from Harvey Mudd College and her PhD in engineering sciences from the Thayer School of Engineering at Dartmouth. She is now a postdoctoral research associate at Boston Children's Hospital with research focused on developing optical imaging methods for pediatric brain research.