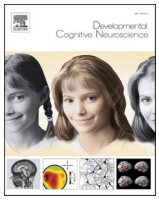
Developmental Cognitive Neuroscience 48 (2021) 100943

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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



Comparing different pre-processing routines for infant fNIRS data

Jessica Gemignani [a,b,*], Judit Gervain [a,b]

[a] Department of Developmental Psychology and Socialisation, University of Padova, Padova, Italy
[b] Integrative Neuroscience and Cognition Center, CNRS & University of Paris, Paris, France

ARTICLE INFO

Keywords:
fNIRS
Infant
Pre-processing
Cognitive developmental neuroscience

ABSTRACT

Functional Near Infrared Spectroscopy (fNIRS) is an important neuroimaging technique in cognitive developmental neuroscience. Nevertheless, there is no general consensus yet about best pre-processing practices. This issue is highly relevant, especially since the development and variability of the infant hemodynamic response (HRF) is not fully known. Systematic comparisons between analysis methods are thus necessary. We investigated the performance of five different pipelines, selected on the basis of a systematic search of the infant NIRS literature, in two experiments. In Experiment 1, we used synthetic data to compare the recovered HRFs with the true HRF and to assess the robustness of each method against increasing levels of noise. In Experiment 2, we analyzed experimental data from a published study, which assessed the neural correlates of artificial grammar processing in newborns. We found that with motion artifact correction (as opposed to rejection) a larger number of trials were retained, but HRF amplitude was often strongly reduced. By contrast, artifact rejection resulted in a high exclusion rate but preserved adequately the characteristics of the HRF. We also found that the performance of all pipelines declined as the noise increased, but significantly less so than if no pre-processing was applied. Finally, we found no difference between running the pre-processing on optical density or concentration change data. These results suggest that pre-processing should thus be optimized as a function of the specific quality issues a give dataset exhibits.

1. Introduction

Functional Near Infrared Spectroscopy (fNIRS) is a non-invasive neuroimaging technique based on the measurement of the optical absorption of cerebral blood (Ferrari and Quaresima, 2012; Scholkmann et al., 2014). Thanks to the different absorption spectra of oxygenated and deoxygenated hemoglobin (HbO and HbR, respectively) in the near-infrared region of the electromagnetic spectrum (650−900 nm), fNIRS measures the relative changes of oxygenation and blood perfusion in the human brain at rest or in response to a specific task.

NIRS is a relatively young technique, but it is gaining increasing recognition in many areas of cognitive neuroscience. One of the most thriving areas of application is developmental neuroscience (Gervain et al., 2011; Aslin et al., 2014). The technique is fully non-invasive, easy-to-use, silent, and well tolerated by even the youngest participants. It doesn't require the use of a tracer substance or a strong magnetic field/pulse. Also, and importantly, infants' skulls and other tissues surrounding the brain are relatively thin, allowing for a deeper penetration of the light into the cortex. At a source-detector distance of 3 cm, NIR light penetrates up to 1–1.5 cm into the cortex in newborns, as opposed to 0.5 cm in adults (Fukui et al., 2003). For all these reasons, NIRS is rapidly becoming one of the imaging techniques of choice in many areas of developmental research, including the study of speech perception and language development (Benavides-Varela and Gervain, 2017; Minagawa-Kawai et al., 2008; Peña et al., 2003), social cognition (Lloyd-Fox et al., 2013, 2009), object perception (Wilcox et al., 2010, 2005), and prediction (Emberson et al., 2017, 2015) in young infants.

As fNIRS is a relatively recent technique, there is no general consensus yet about the best pre-processing practices. This issue is further exacerbated by the fact that the development and variability of the infant hemodynamic response (HRF) is not fully known. Systematic comparisons between analysis methods are thus necessary. The purpose of the current study is, therefore, to investigate the performance of five data pre-processing pipelines, selected from the infant NIRS literature. In Experiment 1, we used synthetic data to compare the recovered HRFs with the true HRF and to assess the robustness of each method against increasing levels of noise. In Experiment 2, we analyzed experimental data from a published study, which assessed the neural correlates of

* Corresponding author at: Via Venezia 8, 35131, Padova, Italy.
E-mail address: jessica.gemignani@unipd.it (J. Gemignani).

https://doi.org/10.1016/j.dcn.2021.100943
Received 30 July 2020; Received in revised form 25 February 2021; Accepted 9 March 2021
Available online 11 March 2021

artificial grammar processing in newborns (Gervain et al., 2012).

1.1. The infant HRF

Infants and toddlers are challenging experimental participants. They cannot comply with instructions, their behavioral repertoire is limited and their attention span is short. When using brain imaging such as NIRS, additional difficulties arise. Infants cannot be instructed to stay motionless, do not tolerate capping well, making adjustments in cap position or combing away hair difficult. These issues result in large and frequent motion artifacts, high frequency noise and other problems in the measured signal. These data quality issues thus need to be taken into account during data processing and analysis. Data quality is often much less good in infant than in adult studies, requiring processing and analysis methods that are specific for infants.

Furthermore and importantly, the characteristics of infant fNIRS data are intrinsically different from adult data. First, despite increasing research efforts, the infant hemodynamic response function (HRF) is not yet fully characterized. In addition to canonical responses (an increase in HbO and a corresponding decrease in HbR), inverted responses (a decrease in HbO and an increase in HbR) are often reported. Such inverted responses have been found across all brain regions (Zimmermann et al., 2012). They tend to appear more frequently between 0 and 3 months of age, but this trend interacts with channel location, the choice of baseline and the nature and complexity of the experimental task and stimuli (Issard and Gervain, 2018). According to a recent review, the morphology of HbR appears to be especially heterogeneous (De Roever et al., 2018). But even canonical infant responses are different from adult responses. Using fMRI, Arichi and colleagues (Arichi et al., 2012) found that the infant HRF typically has smaller amplitude and longer time-to-peak, as well as a significantly deeper undershoot. The delay in peak latency decreases through infancy (Lloyd-Fox et al., 2017). The physiological reasons underlying these differences are likely multi-factorial, but grounded in the substantial developmental changes in brain development and neurovascular coupling in infancy (Roche--Labarbe et al., 2012). The morphology of preterm newborns' HRF correlates significantly with the weight and height at birth, rather than with postmenstrual age (Karen et al., 2019), thus showing individual differences even within the same age group.

These distinctive traits have important implications for data preprocessing and analysis. Even when data quality is similar, the infant HRF is less likely to be detected than its adult equivalent due to its smaller amplitude and longer latency, especially in short presentation blocks or in an event-related design (Aslin et al., 2014). Pre-processing routines for infant data should, therefore, clean the HRF from background noise while preserving as much as possible its amplitude. Furthermore, preprocessing should retain individual HRF shapes, even when those are atypical, since they may provide important insights about the underlying developmental causes, e.g. differential states of brain maturation or different neurocognitive mechanisms triggered by varying stimulus complexity etc.

1.2. The most common pre-processing strategies for infant data

The analysis methods commonly used in the NIRS literature differ in several ways. The most important ones are (i) how motion artifacts are handled, (ii) how physiological noise is filtered and (iii) how preprocessing steps are ordered.

1.2.1. Processing artifacts

When data exhibits motion artifacts, those can either be discarded or corrected. Several motion correction methods have been proposed in the literature and previous studies systematically comparing their effectiveness concluded that wavelet-based filtering (WF, Molavi and Dumont, 2012) often performs best on adults data (Brigadoi et al., 2014; Cooper et al., 2012; Hocke et al., 2018). Studies with fNIRS data from children (Hu et al., 2015) found that the combination of a moving average filter and WF worked even better than WF alone, although the studies noted that WF may reduce the magnitude of the recovered signal. Of particular relevance for the purposes of the present study, for NIRS data acquired from 6−12-month-old infants, WF alone performed better than targeted PCA (tPCA) or the combination of tPCA and WF (Behrendt et al., 2018), whereas for data from infants between 4 and 11 months, the combination of spline interpolation (Scholkmann et al., 2010) and WF performed best with very noisy datasets, while WF alone performed well with moderately noisy datasets (Di Lorenzo et al., 2019). The major advantage of WF is that it retains a large number of experimental trials, whereas its drawbacks are that it reduces the amplitude of the HRF (Brigadoi et al., 2014; Chiarelli et al., 2015) and that it is computationally intensive.

1.2.2. Filtering

A systematic review of filtering methods in adult NIRS studies (Pinti et al., 2019) compared Infinite Impulse Response (IIR) and Finite Impulse Response (FIR) filters of different orders, and found that high order (>500) bandpass FIR filters perform best. When filters were compared in terms of their impact on classification performance in a Brain-Computer Interface study, the *hemodynamic response filter* (Penny et al., 2011) was found to work best (Khan et al., 2020).

1.2.3. Order of pre-processing steps

Pre-processing may be applied to light intensities, optical densities or hemoglobin concentration changes and different processing steps may be applied to different data types. The implications of these choices remain largely unexplored. The only available findings (Pinti et al., 2019) suggest that in terms of the final statistical outcome, there is no difference between applying frequency filtering to optical densities or to hemoglobin concentration changes.

1.3. The current study

Given the specificities of infant NIRS data outlined above, which are particularly marked in newborns and the youngest infants, the question of how different analysis strategies perform in this developmental population is particularly relevant. Yet to date, no study has investigated this systematically.

In this study, we therefore systematically compared several preprocessing pipelines adopted from published infant NIRS studies and show how different data processing options lead to substantially different results in terms of the characteristics of the recovered HRF and data rejection.

Since the use of different processing methods might change the results and thus the conclusions of a study considerably, the choice of processing methods is a particularly timely question. In the last few years, considerable attention and important research efforts have been dedicated to investigating the replicability of studies in psychology and neuroscience (Open Science Collaboration, 2015). Several studies suggest that many published results are not replicable, possibly due in part to cross-laboratory variation in methodological and analysis practices, many of them not explicitly reported or documented in publications (Klein et al., 2014). Diverging analysis practices is an important potential source of variation, possibly compromising replicability. Infant NIRS research is no exception. It is thus crucial to compare and systematically investigate analysis pipelines to better understand how analytic decisions impact results and thus replicability.

To achieve this, we first reviewed the NIRS literature published between 2016 and 2020 with 0−12-month-old infants, and summarized the pre-processing pipelines they used. We then created five prototypical pipelines, which differ from one another parametrically, i.e. in one processing step at a time, and together cover 86 % of the variation among pre-processing methods found in the reviewed literature. The five pipelines are illustrated in Fig. 2 and described in detail in Section

4.2.1.

We compared the impact of these five pipelines as well as of a control pipeline on data quality during the analysis of two datasets. The first was a synthetic dataset that we generated specifically for this study, consisting of synthetic artifacts, in the form of spikes and baseline shifts, physiological confounds and hemodynamic responses, similar to the types of data quality issues found in infant data. This allowed us to compare the HRFs recovered by the five pipelines with the true HRF embedded in the synthesized data, as well as to vary the parameters of the artifacts and the amplitude of the true HRF in a controlled manner in order to assess the robustness of each method against increasing levels of noise. The second dataset was actually measured data from a published study using NIRS to assess the neural correlates of artificial grammar processing in newborns (Gervain et al., 2012).

2. Literature review of pre-processing strategies

2.1. Literature review: selection criteria

A search of fNIRS studies published between 2016 and 2020 was conducted with the goal of reviewing the most common pre-processing methods employed in developmental research. The literature search was performed over the PubMed, Scopus and Web of Science databases using the following criteria:

1 Papers published between 2016 and 2020 (included), i.e. the last 5 years.
2 Studies with infants aged 12 months or below (if the age range of participants spanned the 12 months cut-off, e.g. 10–14 months, the study was included).
3 Peer-reviewed reports of original fNIRS studies. Articles reporting secondary analyses of already published datasets were not included.
4 Studies employing block or event-related stimulus design. Since this study focuses on the recovery of the HRF, functional connectivity and resting state studies were not included.
5 Studies performed using a continuous wave (CW) NIRS instruments.
6 Papers including more than one study were considered the same, if the studies applied the same pre-processing pipeline, separate otherwise.

With these inclusion criteria, 75 studies were selected (a full list is provided in the Supplementary Material). The most important methodological characteristics of the studies are summarized in Fig. 1.

2.2. Selection of pipelines for further analysis

The selected studies varied considerably in their pre-processing strategies. Fig. 2 summarizes this variation. Below, we discuss them following the three pre-processing steps discussed earlier: (i) artifact rejection/correction, (ii) filtering and (iii) order.

2.2.1. Artifacts

Out of the 75 papers, 42 rejected trials contaminated by motion artifacts after identifying them visually (8/42) or using an automatic detection algorithm (33/42), in combination with a visual assessment. Among the latter, many studies reported using either a threshold on the signal amplitude change within a certain time window, or on the standard deviation of the signal, or both. These thresholds were typically fixed for the entire group of participants, but in some studies they were adjusted at the participant level after visual inspection. Among the 33 studies applying motion artifact correction, 14 used WF (Molavi and Dumont, 2012), 10 tPCA (Yücel et al., 2014), 6 spline interpolation (Scholkmann et al., 2010), 2 linear interpolation (Xu et al., 2014) and 1 a combination of several algorithms.

2.2.2. Filtering

The majority of studies performed bandpass filtering, either with a bandpass filter (N = 41) or with the consecutive combination of a low-pass and a high-pass filter (N = 6). Twenty-four studies employed a low-pass filter alone. Three studies did not include information on the frequency filter.

2.2.3. Order

Filtering was applied to hemoglobin concentration changes in 18 studies. In 10 of these, motion artifact detection followed filtering, in 8 it preceded filtering. In 10 instances, motion artifacts detection, trial rejection and filtering were carried out on light intensities, before direct conversion to hemoglobin changes, while in 7 studies these steps were performed on optical densities, before conversion to hemoglobin changes. Among the 33 studies applying artifact correction, the majority (N = 28) did so on optical densities, 2 on light intensities, and 3 on hemoglobin changes.

Based on the above-described trends in the infant NIRS literature and the theoretical questions we sought to answer, we selected the five most

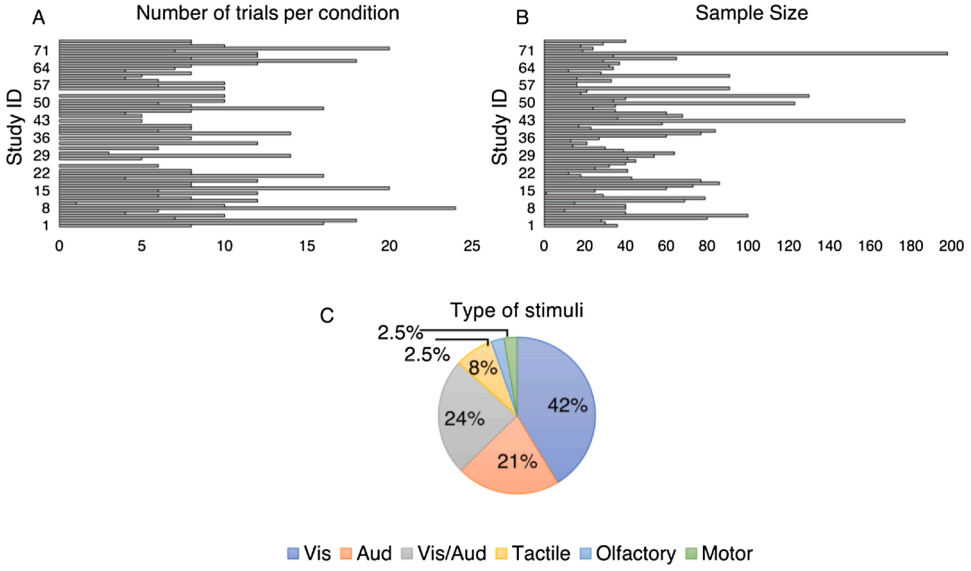


Fig. 1. Summary of the most important methodological properties of the 75 studies identified in the literature search. (A) The average number of trials per condition were 9.2. (B) The average sample size was 46. (C) Proportion of studies using a given stimulus modality.

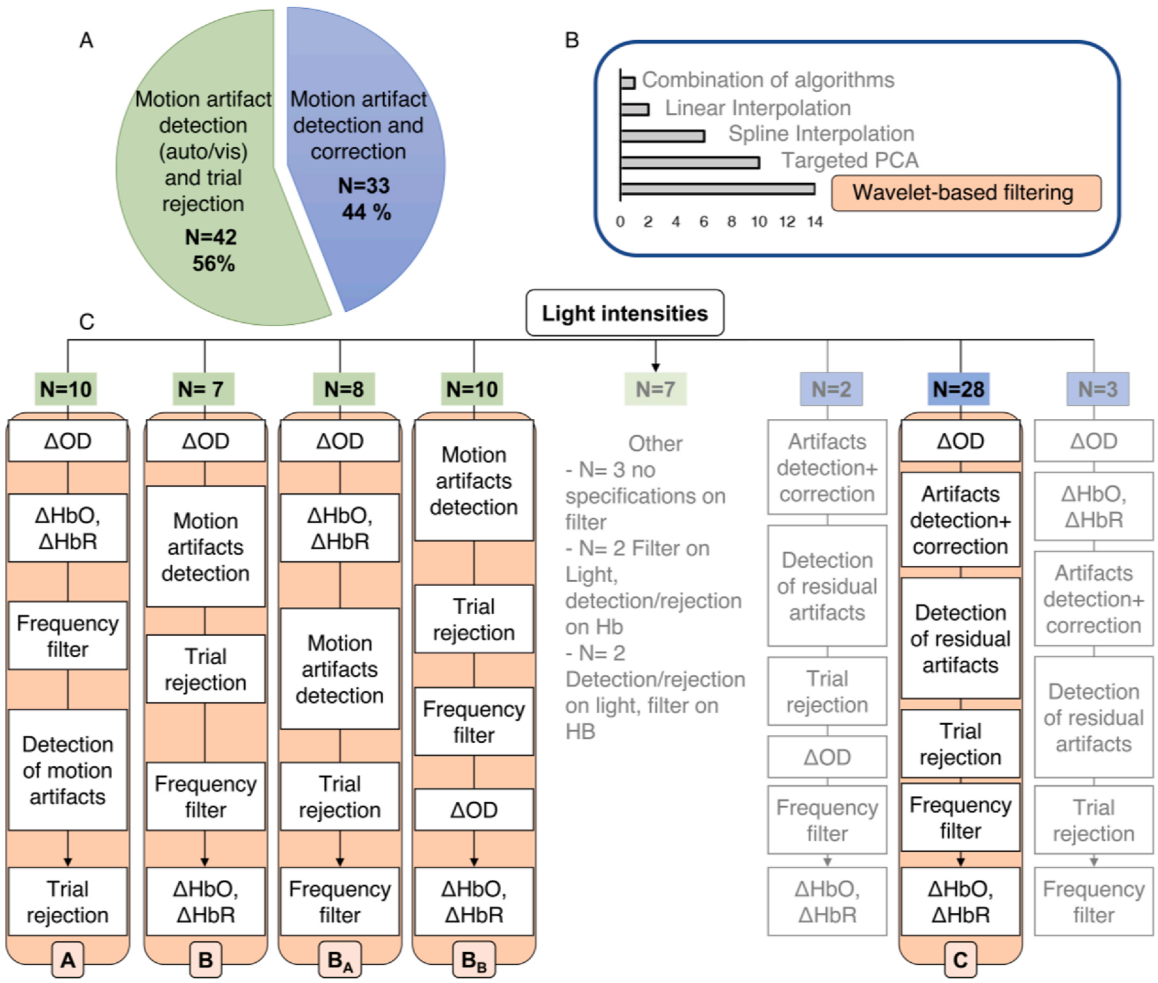


Fig. 2. Summary of the literature review of pre-processing strategies; the pipelines that are selected and compared in this work are those highlighted in orange (A, B, BA, BB and C). Grayed-out pipelines have been employed in very few studies (2 or 3 each) and will not enter the comparison. Among the motion artifacts correction algorithm, WF is selected both because it is the most popular and in light of recent literature. Additionally, a Control pipeline is defined that only applied bandpass filtering and does not control for motion artifacts; the five pipelines are compared to each other and to the Control, that is not used in any study, allowing to have a direct measure of efficacy of each pipeline in dealing with the presence of artifacts.

common pre-processing pipelines (Fig. 2C), labeled A, B, B_A, B_B and C. These pipelines account for 65 studies, i.e. 86 % of the reviewed articles. Importantly, these five pipelines are intended as broad categories, namely they do not represent each cited study in the exact settings and parameters employed, but rather in terms of the three key aspects described above: the way artifacts are handled, frequency filtering and the order of preprocessing steps.

Notably, the five pipelines were chosen in a systematic way, with specific pipeline pairs differing in a single step, making it possible to test the impact of each analysis step on HRF recovery. Thus pipeline C implements artifact correction, while the others implement rejection. The pipelines using trial rejection differ in the order of steps: pipeline A differs from B_A in that filtering and artifact detection are inverted; B, B_A and B_B differ in terms of the data type to which processing is applied (Hb changes, OD and light intensities, respectively). All the pipelines employ a bandpass filter.

The specific settings employed for pipelines A, B, B_A, B_B and C were chosen to best suit the characteristics of the data of this study and are described in Section 4.2.1.

3. Experiment 1

As a first comparison, we assessed the performance of the pipelines on a synthetic dataset with systematically varying parameters for noise and for the HRF.

4. Methods

4.1. Data generation

Synthetic data was generated according to the montage and stimulus design employed in Gervain et al. (2012). In that study, NIRS was acquired in 22 newborns using a montage with 24 channels (Fig. 8). We thus generated a synthetic dataset with 22 "participants", each with 24 time series corresponding to the 24 channels. Like in the original study, the time series comprised 14 trials, each lasting approximately 18 s, and spaced at time intervals of varying duration between 25 and 35 s.

Synthetic data was generated using tools available in the *Brain AnalyzIR Toolbox* for Matlab (Santosa et al., 2018). For each participant, baseline noise was produced by first generating white noise, then imposing temporal correlation on it by employing an autoregressive model of order 30. Different channels were not spatially correlated. Then, to simulate the contribution of heart rate, respiration and Mayer waves to the NIRS signal, the signal amplitude was increased by a factor ranging between 0.01 and 0.03 mM x mm (amounting to about 3–10 % of the total signal change, (Boas et al., 2004)) at frequencies typical of the newborn HRF, namely in the ranges around 1.5 ± 0.2 Hz, 0.25 ± 0.05 Hz and 0.1 ± 0.02 Hz, respectively.

To this "resting state" dataset, we then added HRFs and motion artifacts, simulating functional responses. Twenty different such functional datasets were created by systematically varying the parameters of the HRFs and the artifacts. Motion artifacts were added, in the form of spikes and baseline shifts. Spikes were modelled as a sudden change of voltage ranging between 0.1 and 2 V across the 20 datasets, while baseline shift artifacts were modelled as a random positive or negative change of voltage, also ranging between 0.1 and 2 V. In turn, HRFs had an amplitude value ranging between 0.1 and 0.35 mM x mm for HbO, between -0.05 and -0.175 mM x mm for HbR and an onset-to-peak time of 6 s. HRFs were added to 12 channels, i.e. 50 % of all channels. These will be referred to as "active channels".

Using this procedure, for each participant, channels within the same datasets differed in terms of baseline and physiological noise but shared the same HRF and artifact amplitudes, while the same channel across different datasets shared the same baseline noise, but differed in terms of HRFs and motion artifacts.

The scheme in Fig. 3 describes the simulation steps and shows an example of simulated data. This approach for producing synthetic fNIRS data is similar to the one used in other studies investigating analysis methods (Barker et al., 2013; Gemignani et al., 2018; Huppert, 2016).

4.2. Data analysis

4.2.1. Pre-processing pipelines

The data was pre-processed using the following pipelines. For all the pipelines, the Beer-Lambert Law was applied using the following absorption coefficients ($\mu_{a,}$ mm^{-1} x mM^{-1}): μ_a(HbO, 695 nm) = 0.0955, μ_a(HbO, 830 nm) = 0.232, μ_a(HbR, 695 nm) = 0.451 and μ_a(HbR, 830 nm) = 0.179. The product of the optical pathlength and the differential pathlength factor was set to 1, so that the resulting concentration changes were expressed in mM x mm.

Channel-wise block averages were computed using 5 s before the stimulus onset for the baseline correction, for each block. Grand averages were then calculated by averaging across participants.

4.2.1.1. Pipeline A. Light intensities were first converted to optical densities and to HbO and HbR concentration changes, using the modified Beer-Lambert Law. Subsequently, data was bandpass-filtered between 0.01 and 0.7 Hz, using a *fft* filter. The lower value is defined on the basis of the design, as the expected peak has a frequency of 0.02 Hz (duration of block and inter-block interval ~ 45 s). The higher value allows the suppression of heart rate, sucking on a pacifier and other physiological noise.

Single blocks were rejected if the light intensity reached the saturation value, if the block contained motion artifacts or both. Motion artifacts were defined as concentration changes larger than 0.1 mM x mm over 0.2 s. This procedure was performed on each channel independently. Channels with fewer than 30 % valid blocks per condition were discarded.

For the non-rejected blocks, a baseline was linearly fitted between the means of the 5 s preceding the onset of the block and the 5 s starting 15 s after onset of the block to allow enough time for the HRF to return to baseline.

A conceptually similar pre-processing approach has been employed, with variations in specific settings, in Abboub et al. (2016); Minagawa et al. (2017) and Arimitsu et al. (2018).

4.2.1.2. Pipeline B. Light intensities were converted to optical densities (OD). Motion artifacts were then identified, channel by channel, as changes in signal amplitude of 0.4 or more within a 1 s window, or as 15-fold or larger changes in standard deviation, within 1 s. These parameters were chosen according to the most recent literature (Di Lorenzo et al., 2019; Jackson et al., 2019).

Based on the results of this step, trials were excluded if a motion artifact fell in the stimulation window, defined as starting 2 s before the stimulus onset and ending 10 s after. This operation was also performed channel by channel.

Finally, optical densities were bandpass-filtered between 0.01 and 0.7 Hz with a Butterworth filter of order 3 and concentration changes were computed by applying the Beer-Lambert Law.

A conceptually similar pre-processing approach has been employed, with variations in specific settings, in De Oliveira et al. (2019); Miguel et al. (2019) and Miguel et al. (2020).

4.2.1.3. Pipeline B_A. Light intensities were converted to optical densities and then to relative concentration changes by applying the Beer-Lambert Law.

Then, analogously to Pipeline B, motion artifacts were identified as amplitude changes equal to or greater than 1.3 mM x mm within 1 s, or as 15-fold or greater standard deviation changes within 1 s. Based on the results of this step, trials were excluded if a motion artifact fell in the stimulation window, defined as starting 2 s before the stimulus onset and ending 10 s after.

Finally, concentration changes were bandpass-filtered between 0.01 and 0.7 Hz with a Butterworth filter of order 3.

A conceptually similar pre-processing approach has been employed, with variations in specific settings, in Taga et al. (2018); Hakuno et al. (2020) and Ujiie et al. (2020).

4.2.1.4. Pipeline $B_{B,}$. Light intensities were pre-processed prior to conversion to optical densities and concentration changes. In particular, analogously to Pipelines B and B_A, motion artifacts were identified as amplitude changes equal to or greater than 0.6 V within 1 s, or as 15-fold or greater changes in standard deviation within 1 s. Based on the results of this step, trials were excluded if a motion artifact fell in the stimulation window, defined as starting 2 s before the stimulus onset and ending 10 s after.

Then, data was bandpass-filtered between 0.01 and 0.7 Hz with a Butterworth filter of order 3. Lastly, pre-processed data was converted to optical densities and relative concentration changes, with the modified Beer-Lambert Law.

A conceptually similar pre-processing approach has been employed, with variations in specific settings, in Lloyd-Fox et al. (2017); Mercure et al. (2020) and Van Der Kant et al. (2020).

4.2.1.5. Pipeline C. This pipeline is the same as Pipeline B, except that before detecting motion artifacts, an automatic detection and correction step was performed. To this end, the wavelet-based filtering algorithm proposed by Molavi and Dumont (Molavi and Dumont, 2012) was employed. Briefly, the channel-wise timeseries are decomposed in a series of wavelet detail coefficients. The rationale behind this algorithm is that while coefficients related to the signal of interest are distributed around zero, coefficients representing motion artifacts will lie at the extremes of the distribution. The automatic correction of artifacts is performed by removing these latter coefficients and then reconstructing the signal. The tuning parameter α specifies the boundaries of the distribution beyond which the coefficients are considered outliers and therefore artifacts: coefficients exceeding α times the interquartile range are then removed, and the artifact-free signal is reconstructed using the inverse discrete wavelet transform. Following findings of recent studies (Behrendt et al., 2018; Di Lorenzo et al., 2019), the algorithm was applied using a threshold of $\alpha = 0.5$, and the appropriateness of this choice was confirmed by visual inspection. After automatically detecting and correcting artifacts, the remaining uncorrected motion artifacts were identified using the thresholds described for Pipeline B and the corresponding blocks were rejected.

A conceptually similar pre-processing approach has been employed, with variations in specific settings, in De Klerk et al. (2018); McDonald et al. (2019) and Porto et al. (2020).

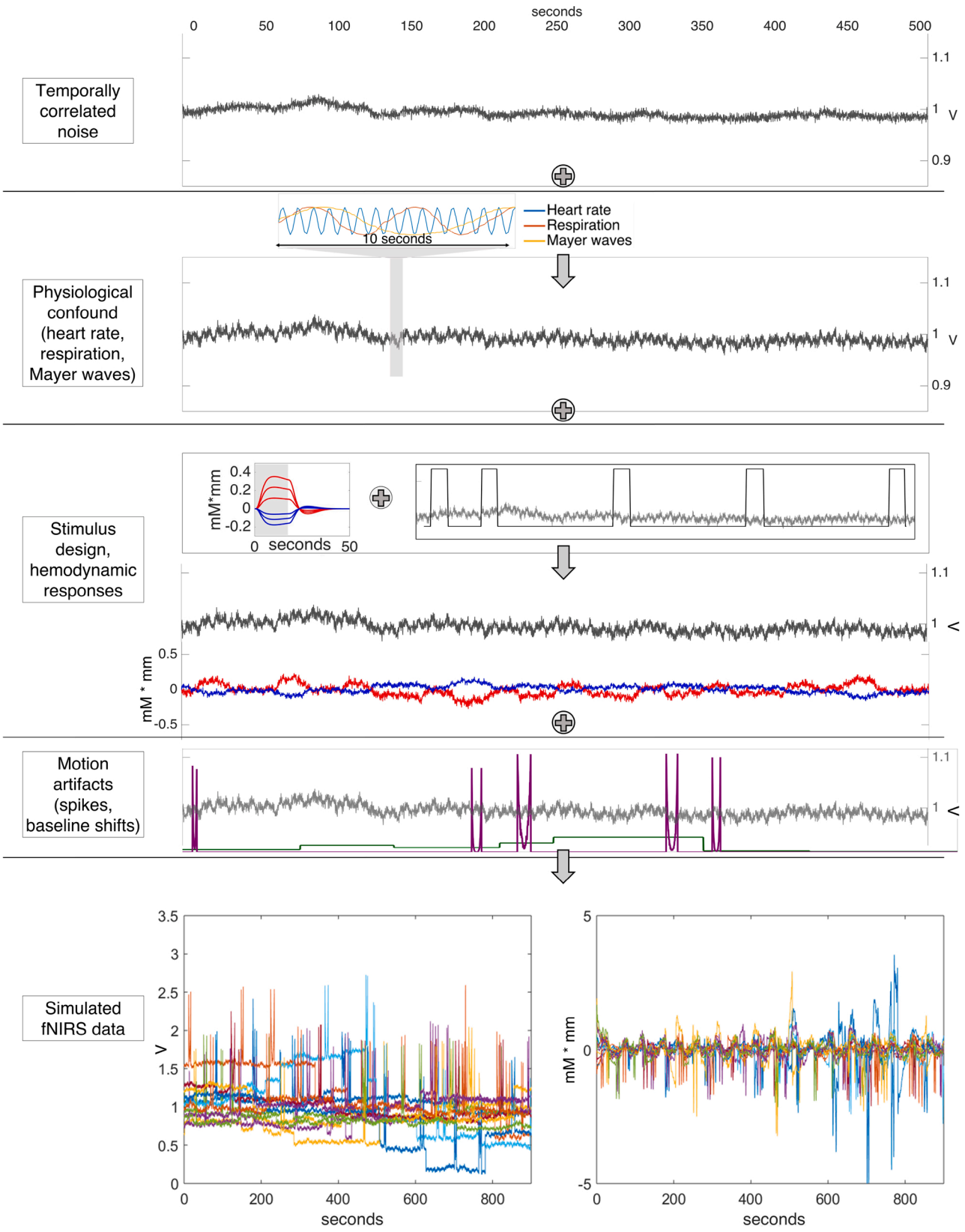


Fig. 3. The scheme describes the workflow that was employed to produce the synthetic dataset. First, temporally correlated noise was generated. Then, the amplitude of the signal was increased at specific frequencies to resemble the contribution of physiological noise. After that, hemodynamic responses were added according to a stimulus design and lastly, motion artifacts were included in form of spikes and shifts. The last panel at the bottom shows an example of a simulated dataset (left: light intensities, right: corresponding concentration changes).

4.2.1.6. Control pipeline. This pipeline was used as the baseline. It consisted of conversion to Hb and filtering using the same parameters as in Pipelines A–C.

4.2.2. Metrics of comparison

The effect of employing the different pre-processing schemes was assessed by comparing several metrics: the root mean square error (RMSE) between the recovered and the true HRF, their correlation, the percentage of included trials and the largest amplitude of the signal in the activation window. The metrics were computed as follows:

RMSE: the RMSE was computed, for each subject and channel, as the square root of the mean squared difference between the block-average of the recovered HRFs and the true HRF.

Correlation: for each subject and channel, the correlation between the average recovered HRFs and the true HRF was evaluated by computing their Pearson correlation coefficient.

Trial inclusion rate: the trial inclusion rate was computed as the proportion of retained trials with respect to the total number (N = 14), averaged across channels.

Amplitude of the signal: The largest positive or negative amplitude of the recovered HRF within the activation window was extracted. The activation window was the time window in which the response in a block was significantly different from the zero baseline, as computed with permutation tests (Abboub et al., 2016; Maris and Oostenveld, 2007).

4.2.3. Statistical analysis

To compare the performance of the pipelines statistically, a mixed effects linear model was fitted to each metric as the dependent variable, with fixed effects for the within-subjects factors Pipeline (A, B, B_A, B_B, C, Control) and Noise (20 levels), as well as their interaction, and random slope for Noise and random intercept for Dataset. Noise in the 20 synthetic datasets was quantified as the coefficient of variation, i.e. the standard deviation of the channel-wise time series divided by their mean amplitudes. We tested all models, starting with the random intercept only model, and adding factors incrementally. We report here, for each metric, the best fitting model, i.e. the one that achieved the lowest Akaike's Information Criterion (AIC). This turned out to be the full model (fixed effects for the within-subjects factors Pipeline and Noise, their interaction, random slope for Noise and random intercept for Dataset) for each of the metrics.

Models were implemented in SPSS v. 25.0 (IBM Corporation, 2017). SPSS outputs an ANOVA-like assessment of the significance of the fixed effects, which we report below. Subsequent pairwise comparisons were adjusted for multiple comparisons with the Bonferroni procedure.

5. Results

The performance of the pipelines on the four metrics is shown in Fig. 4 for HbO. The results of the pairwise comparisons for HbR are reported in the Supplementary Material, as they are highly similar to the HbO results. The recovered HRFs (HbO and HbR) are illustrated in Fig. 5.

5.1. Amplitude

The best fitting (full) mixed effects model yielded a significant main effect of Pipeline (HbO: $F(5, 2153) = 35.78$, $p < 0.001$, HbR: $F(5, 2151) = 44.84$, $p < 0.001$). The pairwise comparisons carrying this main effect are shown in Fig. 4A. Of relevance here is that Pipeline C yielded the lowest amplitude, significantly lower than those recovered by every other pipeline (all *ps*<0.001: mean difference C-A= -0.137, C-B=-0.157, C- $B_{A=}$ -0.156, C-B_B= -0.124) Furthermore, the amplitude of the HRF recovered by Pipeline B_B was significantly lower than the amplitudes of Pipelines B (mean difference = -0.033, $p < 0.001$) and B_A (mean difference = -0.032, $p < 0.001$), as also illustrated in Fig. 6. The distortion

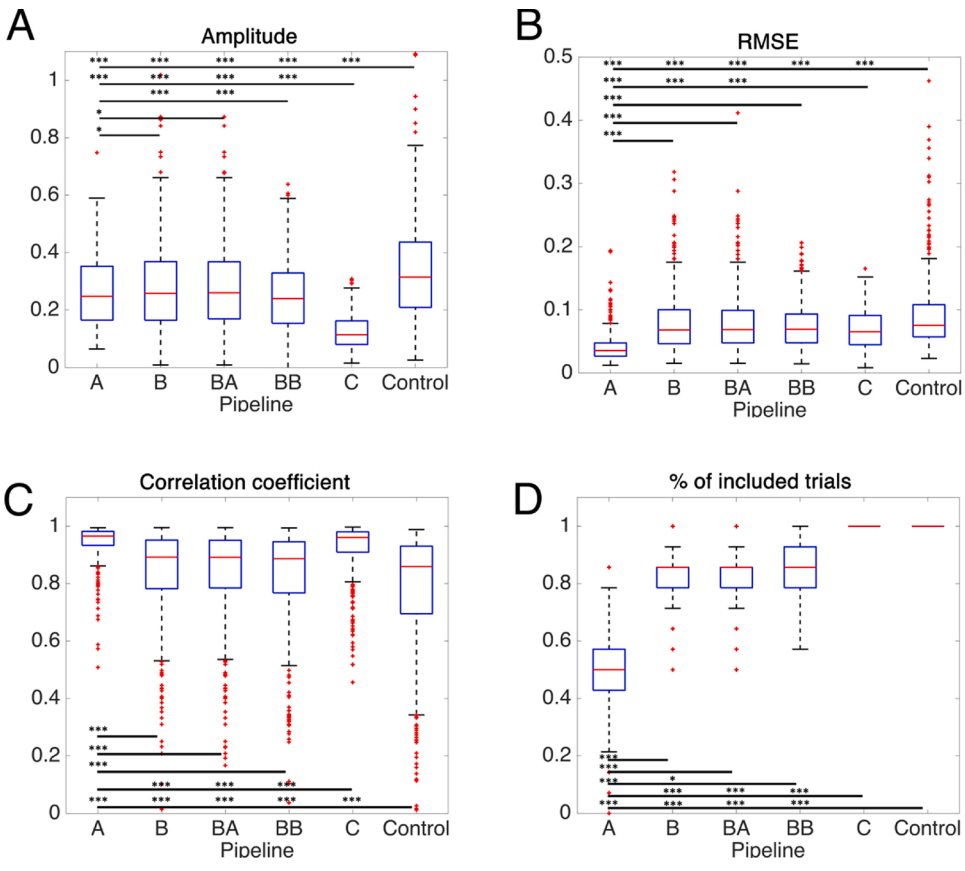


Fig. 4. The amplitude (A), RMSE (B), correlation co-efficient (C) and % included trials (D) achieved by the different processing pipelines averaged across all levels of noise. The red line within each box represents the median value, the two whiskers indicate the first and third quartile and the red asterisks represent outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
Pairwise comparisons are indicated by the lines above the distributions: each line represents the comparison between the level underlying the right extremity of the line and every other level; asterisks indicate the results of the pairwise comparisons: ***, ** and * mark comparisons with $p < 0.001$, 0.01 and 0.05, respectively.

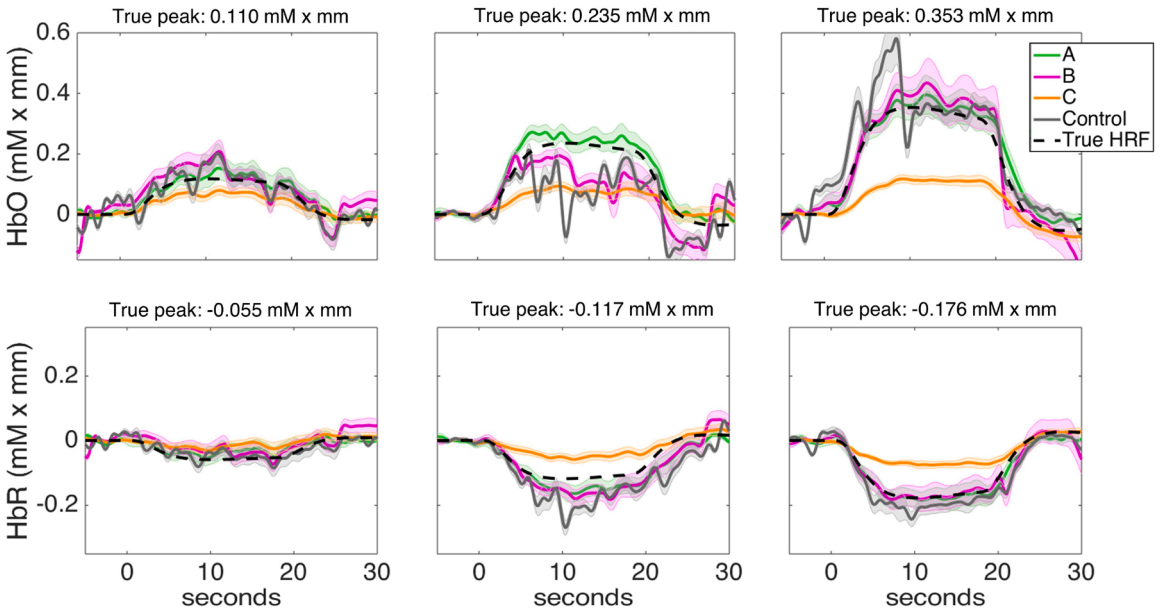


Fig. 5. Three examples of recovered HRFs in randomly chosen "participant" dataset. HRFs are averaged across active channels. Shaded error bars represent the respective standard deviations.

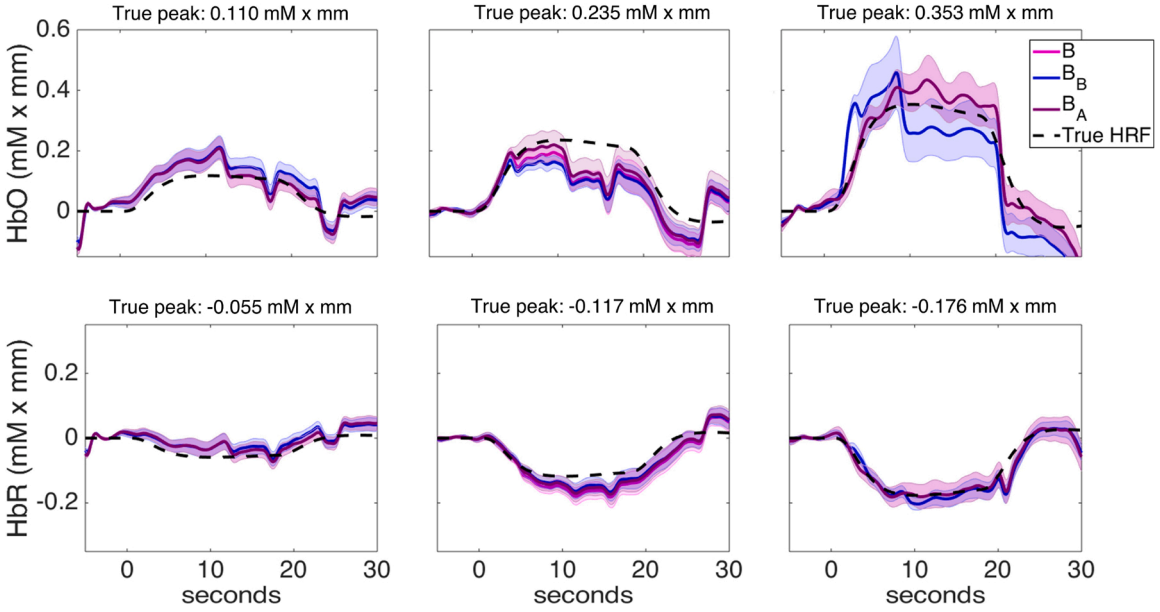


Fig. 6. Three examples of recovered HRFs in randomly chosen "participant" datasets. As in Fig. 5, HRFs are averaged across active channels and shaded error bars represent the respective standard deviations. Note that the averages and standard error bars of Pipelines B (pink) and B_A (violet) largely overlap.

of the HRF shape by Pipeline B_B is particularly evident at higher noise levels (Fig. 7).

The main effect of noise was not significant (HbO: $F(1, 436) = 1.52$, $p > 0.05$, HbR: $F(1,437) = 0.97$), but the Pipeline x Noise interaction was (HbO: $F(5, 2154) = 43.79$, $p < 0.001$, HbR: $F(5, 2152) = 18.3$, $p < 0.001$). This interaction (Fig. 7) was carried by a large increase in amplitude with increasing noise for the Control Pipeline, a small increase for Pipelines B and B_A, and a decrease for Pipelines A, C and B_B.

5.2. RMSE

The best fitting (full) mixed effect model yielded a significant main effect of Pipeline (HbO: $F(5, 2158) = 18.12$, $p < 0.001$, HbR: $F(5,2164) = 27.4$, $p < 0.001$). This was due to lower RMSE for Pipeline A than for the other Pipelines: A–B = -0.037, A-C=-0.27, A–B_A= -0.036, A–B_B= -0.031, all $ps<0.001$. There was also a significant main effect of noise (HbO: $F(1, 435) = 118.9$, $p < 0.001$, HbR: $F(1,439)=184.74$, $p<0.001$). The interaction between Pipeline and Noise was also significant (HbO: $F(5, 2162)=36.79$, $p < 0.001$, HbR: $F(5,2167)=48.97$, $p<0.001$) This was carried by a smaller increase in RMSE with noise for Pipelines A ($\beta = 0.076$), C ($\beta = 0.017$) and B_B ($\beta = 0.068$) than for the other pipelines.

5.3. % Included trials

The best fitting (full) mixed effects model yielded a significant main effect of Pipeline ($F(5, 2190) = 257$, $p < 0.001$). This was carried by all pipelines except pipeline C having a significantly lower inclusion rate than the control pipeline (no rejection) : A-Control= -0.49, B-Control =

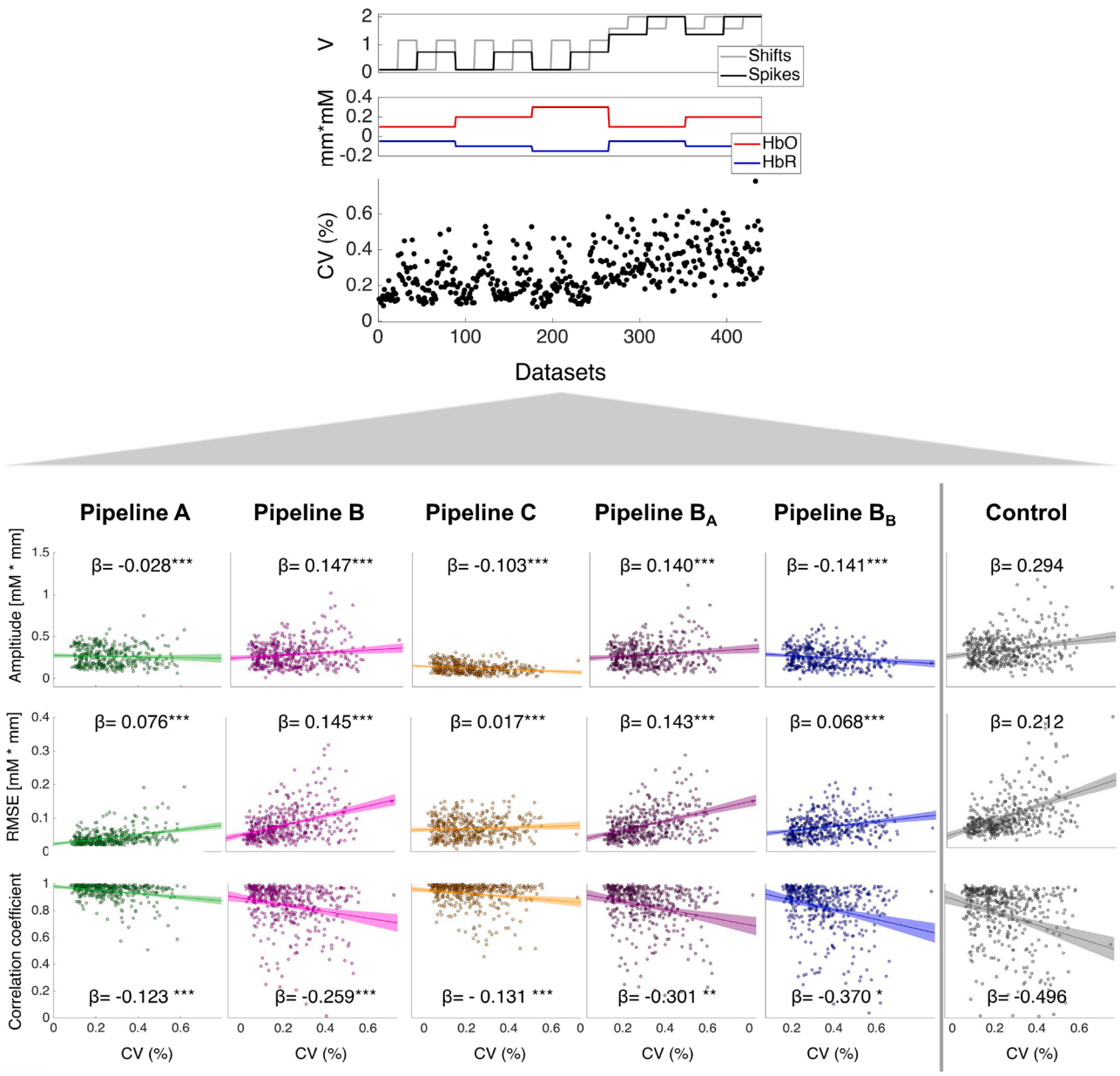


Fig. 7. The top panel shows the parameters of the artifacts used to generate the different datasets, in terms of voltage change. These varying levels of noise are quantified in terms of the coefficient of variation (CV%). The bottom panel shows the performance of each pipeline at different levels of noise for RMSE, Amplitude and the Correlation Coefficient, the metrics for which the Pipeline X Noise interaction was significant. ***, **, * for $p < 0.001$, 0.01 and 0.5, respectively.

-0.181, B_A-Control= -0.176, B_B-Control= -0.164, all *ps*<0.001. Notably, the difference was significant between Pipelines B_B and B (0.018, $p < 0.05$), but not between Pipelines B_B and B_A.. The main effect of Noise ($F(1, 424) = 32.55$, $p < 0.001$) was also significant, as was the Pipeline x Noise interaction ($F(5, 2190) = 67.31$, $p < 0.001$): the decrease in number of retained trials with increasing levels of noise is significant for both Pipeline A (β= -0.54) and B (β = -0.01), but not for the others.

5.4. Correlation between recovered and true HRF

The best fitting (full) mixed effects model yielded a significant main effect of Pipeline (HbO: $F(5, 2156) = 5.13$, $p < 0.001$, HbR: $F(5, 2153) = 6.76$, $p < 0.001$). This was carried by Pipelines A and C yielding the highest correlations (mean value 0.94, SE 0.008 and 0.92, SE 0.008, respectively). The main effect of Noise was also significant ($F(1,437) = 34.62$, $p < 0.001$) as was the interaction between Pipeline and Noise ($F(5, 2158) = 11.68$, $p < 0.001$). This was carried by a smaller decline in correlations for Pipeline A (β = - 0.123) and Pipeline C (β= -0.131) than for B, B_A and B_B (βs = -0.26, -0.30, -0.37, respectively). For all the pipelines the decrease was smaller than applying no pre-processing (Control pipeline, β=-0.496, all *ps* < 0.05).

6. Discussion

The first and most important implication of these results is that pre-processing is a crucial step in the analysis of infant fNIRS data, since every pipeline recovered the HRF better than applying no pre-processing, i.e. the Control Pipeline.

We fine-tuned the characteristics of the synthetic dataset to closely match real infant data: noisy, dense with motion artifacts, and with HRFs that often have small amplitudes, which are difficult to detect. We varied these parameters systematically in order to evaluate the performance of each method as a function of noise and HRF amplitude. We used three metrics, Amplitude, RMSE and the Correlation Coefficient,

related to the shape of the recovered HRF, and one metric, % Included Trial, which assesses the "cost" of recovering the HRF in terms of data inclusion/exclusion.

We found that Pipeline A achieved the best Correlation Coefficient and RMSE among all the pipelines, and it has proven to be the most robust against increasing levels of noise. It thus recovers the HRF particularly faithfully. It does so by using very strict inclusion criteria, i.e. by rejecting the highest number of trials. Such a faithful recovery may be particularly well suited to situations where the targeted effects are small or where the shape of the HRF is itself of interest, e.g. how the shape of the HRF changes over development in a given brain area or cognitive/perceptual domain or whether two populations, e.g. one typical and one atypical group of infants, show differences. Results for HbR point in the same direction (Figure S1).

The HRF recovered by Pipeline C also correlated well with the true HRF and this pipeline maintained its performance quite robustly as the noise increased, but importantly the Amplitude and the RMSE of the recovered HRF were relatively poor. This was likely due to the Wavelet filtering used for motion artifact correction strongly dampening the amplitude of the HRF. This strong reduction of the amplitude of the HRF may be a particularly important issue when testing subtle experimental manipulations with small effect sizes, which is often the case in cognitive developmental studies (e.g. fine-grained perceptual discrimination between two speech sounds, two faces etc.). At the same time, artifact correction makes it possible to retain many more trials, as a result of which Pipeline C had the highest number of retained trials, a useful feature when data is limited, e.g. by participant availability in clinical populations etc.

Pipeline B returned intermediate approximations of the true HRF, but it was more affected by the underlying noise. This can also be observed in the HbR results. Here, only Pipelines A and C yield better correlations than the Control pipeline (Figure S1). and remain robust in the face of increasing levels of noise.

These results reveal a trade-off between the quality of the recovered HRF and inclusion. This balance needs to be considered carefully when processing noisy fNIRS data, such as infant data. On one hand, the correction of motion artifacts is desirable when few trials or few participants are available, e.g. with atypical populations, and thus rejection is not a viable option. On the other hand, if a large number of trials per participant and/or a large sample of participants are available, it is possible for researchers to use stringent quality criteria, by rejecting noisy trials, which in turn allows a better recovery of the HRF. In addition, the overall data quality, e.g. signal-to-noise ratio (SNR), also plays a role, as different pipelines resist noise to different extents. Researchers should thus consider at least the following factors when deciding on the pre-processing steps: (i) whether the exact characterization of the shape of the HRF is relevant; (ii) what the expected effect size is; (iii) the number of trials per participant and the number of participants, and (iv) the data quality (SNR, the quantity and type of artifacts etc.). Similar trade-offs between data inclusion and quality have also been documented in behavioral infant data, such as looking time measures (ManyBabies Consortium, Frank et al. (2020)), suggesting that this compromise is inherent in noisy datasets in general.

We also tested the impact of the order of pre-processing steps by comparing the same pre-processing applied to optical densities (Pipeline B), to Hb concentrations (Pipeline B_A) or to raw light intensities (Pipeline B_B). We found little difference between using optical densities or HbO/HbR concentrations. This result is in line with the findings reported in Pinti et al. (2019), who also found that the two approaches did not differ in their statistical outcomes (in a General Linear Model). Our analysis shows that this is because applying the pre-processing to optical densities or to hemoglobin concentration changes does not produce significant differences in the size and shape of the recovered HRF. Interestingly, however, applying the pre-processing steps on light intensities (B_B) resulted in lower amplitudes. The recovered HRF was also more irregular in shape. We speculate that the filter specifications that work well with optical densities and concentration changes, such as a 3rd order bandpass Butterworth filter, might not have been equally appropriate for use with light intensities. As it has been suggested by other studies (Pinti et al., 2019), future work should investigate the optimal filter parameters specifically suited for infants data.

7. Experiment 2

Applying the pipelines to synthetic data allowed us to explore their performance in a systematic manner. In Experiment 2, we now test them on real data in order to explore the ecological validity and generalizability of the findings from Experiment 1.

8. Methods

8.1. Data

The data was obtained from Gervain et al. (2012), Experiment 1. The purpose of the study was to assess whether the newborn brain discriminates random sequences from repetition-based regularities in speech stimuli. Here, we briefly summarize the methodological details regarding the participants, the stimuli and the experimental procedures.

8.1.1. Participants

Twenty-two healthy, full-term neonates (13 boys, 9 girls; mean age = 1.14 days, range = 0–3 days; Apgar score $\geq$ 8) born in the Vancouver area participated in the experiment. Data from 13 additional infants were collected but excluded from the data analysis, as they (1) failed to complete the experiment because of fussiness and crying (11 infants) or (2) provided poor quality data because of large motion artifacts or thick hair (2 infants). All parents gave informed consent before participation. The ethics boards of the University of British Columbia and BC Women's Hospital, where the experiments took place, granted permission.

8.1.2. Stimuli and procedure

Infants were tested with a HITACHI ETG-4000 NIRS machine (source-detector separation: 3 cm; two continuous wavelengths of 695 and 830 nm; sampling rate: 10 Hz) using 24 channels over the bilateral temporal, parietal and frontal areas (Fig. 8B). Auditory stimuli were presented according to a block design, in which each block (trial) included 10 trisyllabic words characterized by sequence-initial repetitions ("AAB", e.g. "babamu", "nanape") or by random sequences ("ABC", e.g. "mubage", "penaku"). The stimuli were synthesized to have monotonous pitch (200 Hz) and equal syllable durations (270 ms). A detailed description of the stimuli can be found in (Gervain et al., 2012).

A total of 14 blocks for each condition was presented; blocks lasted approximately 18 s, were spaced by time intervals of varying duration (25−35 s) and were presented in a randomized order (Fig. 8A).

8.2. Data analysis

8.2.1. Pre-processing pipelines

The pre-processing pipelines were identical to those used in Experiment 1.

8.2.2. Metrics of comparison

The pipelines were compared in terms of percentage of included trials and signal amplitude, as described in Experiment 1. Since the true HRF was not known, the RMSE and the Correlation Coefficients could not be computed. We also added a new measure, the standard deviation of the HRF in order to assess the noise and variability in the recovered signal.

8.2.3. Statistical analysis

Statistical analyses for signal amplitude and the standard deviation were performed on the channels that were reported to have a significant

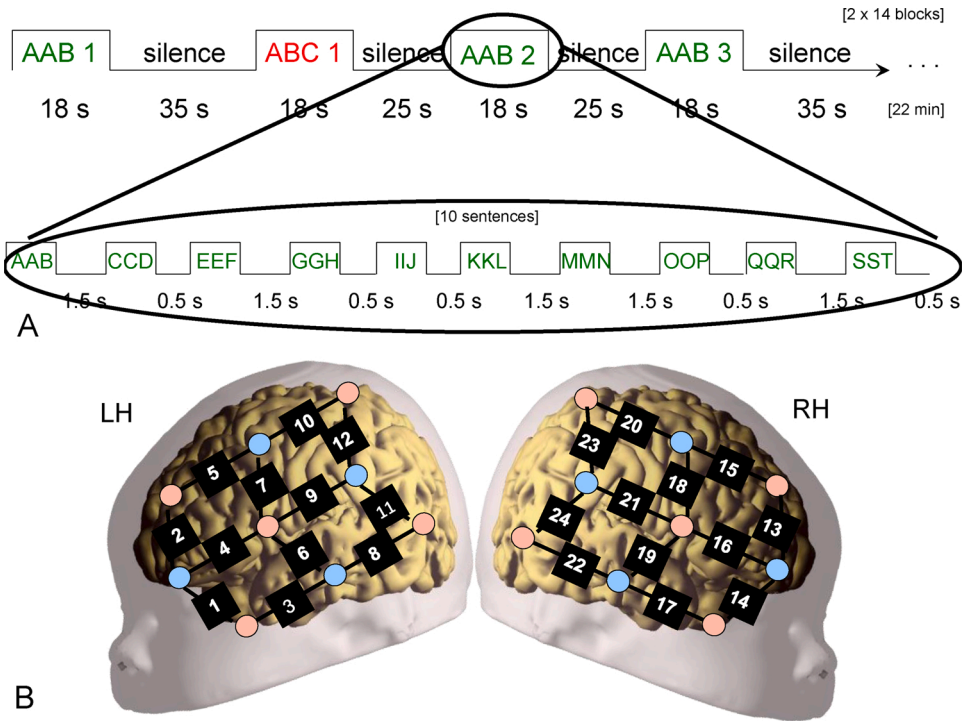


Fig. 8. The design of the experiment (figure adapted from Gervain et al., 2012).

response to the AAB condition in Gervain et al. (2012), i.e. they showed a detectable hemodynamic response (Gervain et al., 2012), in a time window defined using a permutation test with 100 iterations comparing the AAB responses to a zero baseline. The channels were 3, 4, 6 and 15 (Fig. 8). Of relevance here is the fact that these channels have varying levels of noise, with channel 4 being particularly noisy (Fig. 9, insets in the top and middle row plots) Specifically, the standard deviation of the signal averaged across blocks and participants, is 0.3416 for channel 4, as opposed to 0.17, 0.18 and 0.16 for channels 3, 6 and 15, respectively.

Linear mixed effects models were fitted to HRF amplitude and standard deviation, separately, as the dependent variable, with fixed effects for Pipeline (A, B, B_A, B_B, C) and Channel (3, 4, 6, 15) and their interaction, as well as a random slope for channel and a random intercept for participant. This model was selected based on the minimization of the AIC criterion, analogously to Experiment 1.

The percentage of trial inclusion was averaged across all channels and across the two conditions, and a linear mixed model was fitted to it with a fixed effect for Pipeline and a random intercept for participant, which was the best fitting model based on AIC.

Models were implemented in SPSS v. 25.0 (IBM Corporation, 2017) and subsequent pairwise comparisons were adjusted for multiple comparisons with the Bonferroni procedure.

9. Results

The recovered HRFs for HbO and the corresponding amplitude and standard deviation values are shown in Fig. 9. The results for HbR are highly similar and are, therefore, shown in the Supplementary Material. But we report the statistics for both Hb species here.

The mixed effects model over amplitude values yielded a significant effect for Pipeline (HbO: $F(5, 368) = 24.9$, $p < 0.001$, HbR: $F(5, 366) = 13.86$, $p < 0.001$), for Channel (HbO: $F(3, 76) = 2.78$, $p = 0.046$, HbR ns), as well as for their interaction (HbO: $F(15, 368) = 3.55$, $p < 0.001$, HbR: $F(15, 366) = 3.42$, $p < 0.001$). The main effect of Pipeline was mainly due to Pipeline C producing lower HRF amplitudes than the other pipelines (mean differences C-A −0.067, C-B −0.059, C-B_A −0.084, C-B_B −0.11, all $ps < 0.001$). The main effect of Channel was carried by Channel 15, characterized by a significantly lower amplitude than the others: mean difference between Channel 15 and Channel 3−0.059, $p < 0.05$, Channel 15−4 −0.035, $p < 0.05$; Channel 15−6: −0.054, $p < 0.001$. The significant Pipeline x Channel interaction (post hoc test results are shown in Fig. 9) was mainly attributable to the amplitude obtained by Pipeline C being significantly lower than the amplitudes of the other pipelines to different extents in the different channels. The amplitude recovered by Pipeline B was not different from that recovered by Pipeline B_A in any of the channels, while B_B had significantly higher amplitudes than the other pipelines in channel 4, which was particularly noisy.

The mixed effects model over the standard deviation of the HRF yielded a significant effect for Pipeline (HbO: $F(5, 417) = 23.35$, $p < 0.001$, HbR: $F(5, 411) = 22.88$, $p < 0.001$), for Channel (HbO ns, HbR: $F(3, 120) = 3.47$, $p < 0.05$) and for their interaction (HbO: $F(15, 417) = 3.42$, $p < 0.001$, HbR: $F(15, 411) = 4.26$, $p < 0.001$). The main effect of Pipeline was driven by the three Pipelines B (B-B_B-B_A) yielding significantly greater variability in the HRF than the other pipelines. Post-hoc tests revealed no significant differences between Pipelines B, B_A and B_B, but all three had higher standard deviations than Pipelines A (B-A 0.025, $p < 0.01$; B_A-A 0.029, $p < 0.001$; B_B-A 0.028, $p < 0.001$) and C (B-C 0.043, $p < 0.001$; B_A-C 0.046, $p < 0.001$; B_B-C 0.046, $p < 0.001$). The main effect of Channel was not carried by any significant pairwise comparisons in the HbO timetraces, while in HbR Channel 3 displays a significantly lower standard deviation than Channel 4 (mean difference 3-4 = −0.011, $p < 0.05$). Lastly, the significant Pipeline x Channel interaction is mainly due to Channel 4 yielding significant differences between Pipelines more than the other channels (post-hocs are shown in Fig. 9).

The trial inclusion rates are shown in Fig. 10. The effect of pipeline was significant ($F(5, 105) = 135.9$, $p < 0.001$), with all post hoc pairwise comparisons being statistically significant, except between the Control pipeline and Pipeline C, as well as between Pipelines B, B_A and B_B.

10. Discussion

The analysis of experimental data provided similar results to those

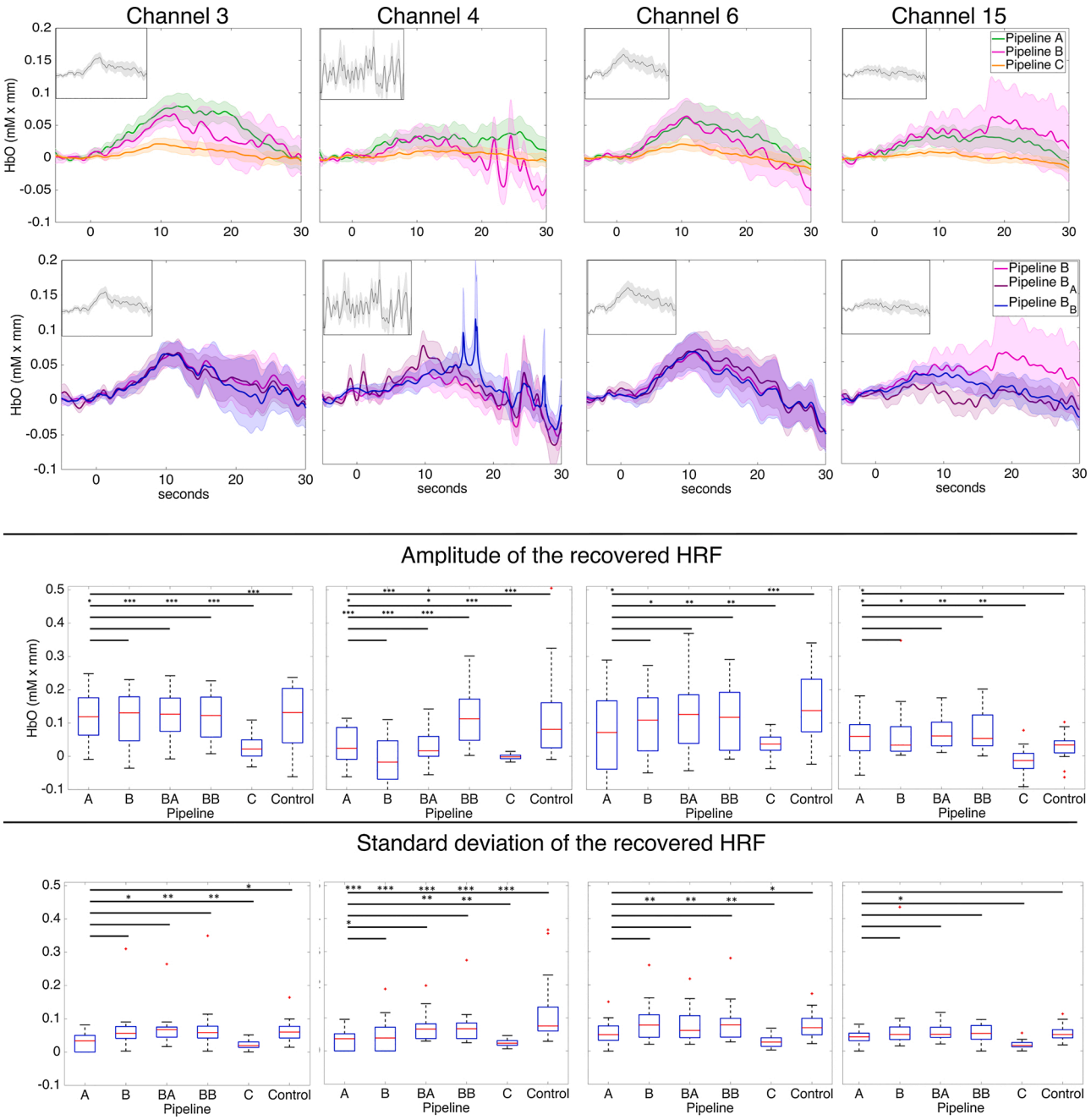


Fig. 9. (First line) Grand averages of the AAB condition obtained through Pipelines A, B and C across all subjects for channels 3, 4, 6 and 15. The inset plots show the corresponding grand averages for the Control Pipeline. **(Second line)** Grand averages of the same channels obtained when using Pipelines B, B_A and B_B. **(Third line)** HRF amplitudes recovered in the same channels by the different pipelines (**Fourth line**) Standard deviations of the HRFs in the same channels and pipelines.

found for the synthetic data. Wavelet-based artifact correction retained a large amount of data, at the cost of yielding a low amplitude HRF. Conversely, the pipelines that involved trial exclusion, i.e. Pipelines A, B, B_A and B_B, showed almost no differences in amplitude, and yielded a more discernible HRF.

There was, however, a difference between these pipelines in how variable (noisy) the recovered HRFs were, with Pipelines A and C yielding the cleanest timeseries, and Pipelines B, B_A and especially B_B yielding HRFs with large standard deviations, particularly in the noisiest channel (Fig. 9, Channel 4), for both HbO and HbR. This confirms what we had already observed in Experiment 1: Pipelines A and C are robust in the face of noise, while Pipeline B_B performs poorly under noisy conditions, likely because its filter specifications are sub-optimal for light intensities.

Pipelines also varied in the amount of trials they excluded: A discarded the largest number; higher rates are achieved by the other pipelines. This confirms the inclusion-quality trade-off, also observed in Experiment 1.

11. General discussion and recommendations

The use of fNIRS has been growing rapidly in many areas of neuroscience research. In particular, one of the most prolific areas of application is the field of infant research (Gervain et al., 2011; Minagawa-Kawai et al., 2008). Nevertheless, being a relatively recent neuroimaging technique, standardized testing and analysis procedures are still lacking. This issue is especially pressing for infant studies, since the development of the hemodynamic response function (HRF) in

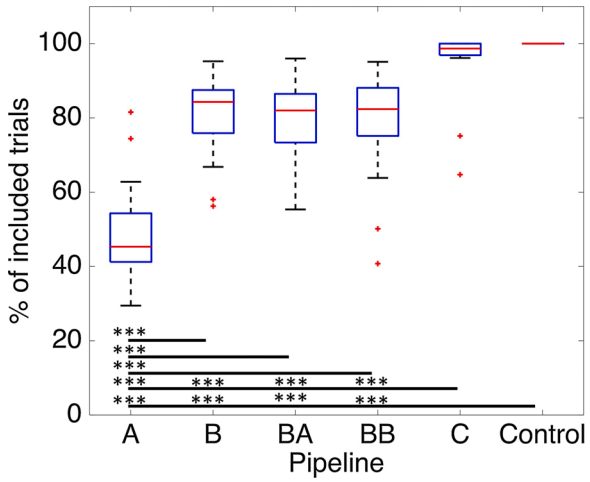


Fig. 10. Trial inclusion rates across pipelines averaged across channels and across the two conditions AAB and ABC. The black lines indicate the results of the pairwise comparisons, and the asterisks mark comparisons with $p < 0.001$.

infants is not fully known and exhibits great variability across subjects, experimental designs, age and brain areas.

Importantly, infants are non-compliant participants with little control over their head motion and have short attention spans, yielding noisy and artifacted data. For all these reasons, infant data is inherently variable and data quality may often be compromised.

Data analysis therefore plays a crucial role in how well the HRF can be recovered. But many different data processing pipelines are used and reported in literature.

The goal of this work was, therefore, to select a few representative pre-processing workflows from the existing infant NIRS literature and assess their performance. We performed a literature search, which identified 75 studies carried out with infants aged 12 months or younger, published between 2016 and 2020. Based on this, we defined five pre-processing pipelines that accounted for 86 % of the reviewed literature, and we compared their performance qualitatively and quantitatively on the same synthetic (Experiment 1) and real (Experiment 2) datasets.

In both experiments, we observed that the pre-processing pipelines all performed better than the control pipeline, which included only filtering. Furthermore and more interestingly, we observed a trade-off between the quality of the recovered HRF and data inclusion. The percentage of included trials was highest when correcting motion artifacts with a wavelet-based filtering. The possibility of retaining a large quantity of data is the greatest asset of motion artifacts correction techniques. This is particularly relevant for infant studies or for studies with atypical, clinical or otherwise hard-to-recruit populations, where experimental time is limited, and it is impossible to fully control the participants' behavior.

However, the amplitude of the HRF recovered by correcting motion artifacts was found to be much lower than that of the true HRF, while the pipelines that did not apply this correction recovered the HRF with amplitudes closer to the true ones, in Experiment 1, and very similar to one another, in Experiment 2. These pipelines are thus more suitable for studies in which the shape of the HRF is relevant, where subtle differences between conditions or small effect sizes are to be expected and where appropriate amounts of data can be collected.

In light of these results, we argue that researchers should consider a number of factors when selecting the pre-processing steps. First, the data quality/inclusion trade-off needs to be taken into account. Researchers need to consider whether a more accurate identification of the HRF or larger data inclusion is more desirable to best answer their research question.

Second and relatedly, researchers need to take into account the quality of the raw data. We found that the performance of all pipelines deteriorated with increasing levels of noise, but pipelines A and C were considerably less impacted. When data quality is poor, the pipeline should be chosen accordingly.

Third, the automatic correction of motion artifacts should be used cautiously. Several settings should be tested and verified. Since the amplitude of infant data is generally lower than that of adult data, and differences between experimental conditions may be small, decreasing it further may be detrimental, and could result in false negatives, reduced effect sizes and an underestimation of the hemodynamic response.

Fourth, both experiments suggest that applying pre-processing to optical densities or concentration changes achieved overall very similar results, while applying it to light intensities yielded more irregular responses, especially when dealing with considerable underlying noise. We speculate that the 3rd order Butterworth filter is likely not well suited for light intensity data, and more work is necessary to identify the most appropriate filter design for this case.

The findings of this work highlight how pre-processing choices have an impact on the shape and the amplitude of the recovered hemodynamic response, and therefore on the conclusions of a study. Since neuroscience and psychology are increasingly concerned with the (non-) replicability of experimental findings, well-motivated analysis choices become central in ensuring the robustness of NIRS studies. The transparent and explicit reporting of these choices is of utmost importance.

While a one-size-fits-all approach is not feasible, as factors such as the number of available trials, the behavioral state of the participants, their age, the nature of the stimuli etc. all contribute to determine the best analysis choices, we nevertheless, provide some guidelines that can help researchers in designing a suitable pre-processing strategy.

- In studies with atypical, clinical or otherwise hard-to-recruit populations, where participant availability and experimental time are strongly limited, and it is particularly challenging to control the participants' behavior, it is advisable to attempt the correction of motion artifacts.
- Whenever it is possible to test for extended periods of time, e.g. with sleeping newborns, and a large number of trials can be obtained, artifact rejection can be applied in order to most faithfully recover the HRF, especially its amplitude.
- If small effect sizes are to be expected based on literature, we recommend rejecting bad quality trials: this approach yields the highest amplitude and lowest standard deviation within blocks, contributing to a higher effect size.
- When it is important that two conditions be maximally discriminable (e.g. in clinical applications, when a pathological state needs to be distinguished from a typical one), it is recommendable to reject bad quality trials.
- When the study seeks to investigate or characterize the shape of the HRF itself in a given population, brain area, age range or task, it is reasonable to reject bad quality trials in order to best preserve the shape and amplitude of the HRF.

12. Conclusions

Although fNIRS is widely employed in the field of cognitive developmental research, a consensus on a common pre-processing workflow is lacking. In this work we selected five commonly employed pre-processing pipelines from recent infant fNIRS literature. The pre-processing stage has a critical impact on the shape of the recovered HRF and therefore, ultimately, on the interpretation of the results. By using both experimental and synthetic data, we demonstrated the strengths and limitations of different pre-processing choices; in particular, we highlighted that automatic correction of motion artifacts allows to retain the vast majority of noisy trials, but also produces a reduction in the recovered response amplitude. By contrast, employing strict criteria for trial inclusion results in a large exclusion rate, which is not

always feasible, but better preserves the characteristics and amplitude of the HRF. Finally, the same pre-processing steps may invariably be performed on optical densities or concentration changes, but are not suited to raw light intensities.

Author contributions

JG and JG developed the study concept. Data generation and analysis was performed by J Gemignani. JG and JG wrote the manuscript. All authors contributed to the study design and approved the final version of the manuscript for submission.

Research data for this article

The data that support the findings of this study are available from the authors upon reasonable request.

Funding

This work was supported by the ERC Consolidator Grant "BabyRhythm 773202" awarded to Judit Gervain.

Declaration of Competing Interest

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

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.dcn.2021.100943.

References

Abboub, N., Nazzi, T., Gervain, J., 2016. Prosodic grouping at birth. Brain Lang. 162, 46–59. https://doi.org/10.1016/j.bandl.2016.08.002.

Arichi, T., Fagiolo, G., Varela, M., Melendez-Calderon, A., Allievi, A., Merchant, N., Tusor, N., Counsell, S.J., Burdet, E., Beckmann, C.F., Edwards, A.D., 2012. Development of BOLD signal hemodynamic responses in the human brain. Neuroimage 63, 663–673. https://doi.org/10.1016/j.neuroimage.2012.06.054.

Arimitsu, T., Minagawa, Y., Yagihashi, T.O., Uchida, M., Matsuzaki, A., Ikeda, K., Takahashi, T., 2018. The cerebral hemodynamic response to phonetic changes of speech in preterm and term infants: The impact of postmenstrual age. Neuroimage Clin. 19, 599–606. https://doi.org/10.1016/j.nicl.2018.05.005.

Aslin, R.N., Shukla, M., Emberson, L.L., 2014. Hemodynamic correlates of cognition in human infants. Annu. Rev. Psychol. 66, 349–379. https://doi.org/10.1146/annurev-psych-010213-115108.

Barker, J.W., Aarabi, A., Huppert, T.J., 2013. Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. Biomed. Opt. Express 4, 1366. https://doi.org/10.1364/BOE.4.001366.

Behrendt, H.F., Firk, C., Nelson, C.A., Perdue, K.L., 2018. Motion correction for infant functional near-infrared spectroscopy with an application to live interaction data. Neurophotonics 5, 1. https://doi.org/10.1117/1.nph.5.1.015004.

Benavides-Varela, S., Gervain, J., 2017. Learning word order at birth: a NIRS study. Dev. Cogn. Neurosci. 25, 198–208. https://doi.org/10.1016/j.dcn.2017.03.003.

Boas, D.A., Dale, A.M., Franceschini, M.A., 2004. Diffuse optical imaging of brain activation: approaches to optimizing image sensitivity, resolution, and accuracy. NeuroImage. https://doi.org/10.1109/UCC.2014.70.

Brigadoi, S., Ceccherini, L., Cutini, S., Scarpa, F., Scatturin, P., Selb, J., Gagnon, L., Boas, D.A., Cooper, R.J., 2014. Motion artifacts in functional near-infrared spectroscopy: a comparison of motion correction techniques applied to real cognitive data. Neuroimage 85, 181–191. https://doi.org/10.1016/j.neuroimage.2013.04.082.

Chiarelli, A.M., Maclin, E.L., Fabiani, M., Gratton, G., 2015. A kurtosis-based wavelet algorithm for motion artifact correction of fNIRS data. Neuroimage 112, 128–137. https://doi.org/10.1016/j.neuroimage.2015.02.057.

Cooper, R.J., Selb, J., Gagnon, L., Phillip, D., Schytz, H.W., Iversen, H.K., Ashina, M., Boas, D.A., 2012. A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. Front. Neurosci. 6, 1–10. https://doi.org/10.3389/fnins.2012.00147.

De Klerk, C.C.J.M., Hamilton, A.Fd.C., Southgate, V., 2018. Eye contact modulates facial mimicry in 4-month-old infants: an EMG and fNIRS study. Cortex 106, 93–103. https://doi.org/10.1016/j.cortex.2018.05.002.

De Oliveira, S.R., Machado, A.C.C.P., de Paula, J.J., Novi, S.L., Mesquita, R.C., de Miranda, D.M., Bouzada, M.C.F., 2019. Changes of functional response in sensorimotor cortex of preterm and full-term infants during the first year: an fNIRS study. Early Hum. Dev. 133, 23–28. https://doi.org/10.1016/j.earlhumdev.2019.04.007.

De Roever, I., Bale, G., Mitra, S., Meek, J., Robertson, N.J., Tachtsidis, I., 2018. Investigation of the pattern of the hemodynamic response as measured by functional near-infrared spectroscopy (fNIRS) studies in newborns, less than a month old: a systematic review. Front. Hum. Neurosci. 12 https://doi.org/10.3389/fnhum.2018.00371.

Di Lorenzo, R., Pirazzoli, L., Blasi, A., Bulgarelli, C., Hakuno, Y., Minagawa, Y., Brigadoi, S., 2019. Recommendations for motion correction of infant fNIRS data applicable to multiple data sets and acquisition systems. Neuroimage 200, 511–527. https://doi.org/10.1016/j.neuroimage.2019.06.056.

Emberson, L.L., Richards, J.E., Aslin, R.N., 2015. Top-down modulation in the infant brain: learning-induced expectations rapidly affect the sensory cortex at 6 months. Proc. Natl. Acad. Sci. U. S. A. 112, 9585–9590. https://doi.org/10.1073/pnas.1510343112.

Emberson, L.L., Zinszer, B.D., Raizada, R.D.S., Aslin, R.N., 2017. Decoding the infant mind: multivariate pattern analysis (MVPA) using fNIRS. PLoS One 12, e0172500. https://doi.org/10.1371/journal.pone.0172500.

Ferrari, M., Quaresima, V., 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. Neuroimage 63, 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049.

Frank, M.C., Alcock, K.J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S., Bergelson, E., Bergmann, C., Black, A.K., Blything, R., Böhland, M.P., Bolitho, P., Borovsky, A., Brady, S.M., Braun, B., Brown, A., Byers-Heinlein, K., Campbell, L.E., Cashon, C., Choi, M., Christodoulou, J., Cirelli, L.K., Conte, S., Cordes, S., Cox, C., Cristia, A., Cusack, R., Davies, C., de Klerk, M., Delle Luche, C., Ruiter, Lde, Dinakar, D., Dixon, K.C., Durier, V., Durrant, S., Fennell, C., Ferguson, B., Ferry, A., Fikkert, P., Flanagan, T., Floccia, C., Foley, M., Fritzsche, T., Frost, R.L.A., Gampe, A., Gervain, J., Gonzalez-Gomez, N., Gupta, A., Hahn, L.E., Kiley Hamlin, J., Hannon, E.E., Havron, N., Hay, J., Hernik, M., Höhle, B., Houston, D.M., Howard, L. H., Ishikawa, M., Itakura, S., Jackson, I., Jakobsen, K.V., Jarto, M., Johnson, S.P., Junge, C., Karadag, D., Kartushina, N., Kellier, D.J., Keren-Portnoy, T., Klassen, K., Kline, M., Ko, E.-S., Kominsky, J.F., Kosie, J.E., Kragness, H.E., Krieger, A.A.R., Krieger, F., Lany, J., Lazo, R.J., Lee, M., Leservoisier, C., Levelt, C., Lew-Williams, C., Lippold, M., Liszkowski, U., Liu, L., Luke, S.G., Lundwall, R.A., Macchi Cassia, V., Mani, N., Marino, C., Martin, A., Mastroberardino, M., Mateu, V., Mayor, J., Menn, K., Michel, C., Moriguchi, Y., Morris, B., Nave, K.M., Nazzi, T., Noble, C., Novack, M.A., Olesen, N.M., John Orena, A., Ota, M., Panneton, R., Esfahani, S.P., Paulus, M., Pletti, C., Polka, L., Potter, C., Rabagliati, H., Ramachandran, S., Rennels, J.L., Reynolds, G.D., Roth, K.C., Rothwell, C., Rubez, D., Ryjova, Y., Saffran, J., Sato, A., Savelkouls, S., Schachner, A., Schafer, G., Schreiner, M.S., Seidl, A., Shukla, M., Simpson, E.A., Singh, L., Skarabela, B., Soley, G., Sundara, M., Theakston, A., Thompson, A., Trainor, L.J., Trehub, S.E., Trøan, A.S., Tsui, A.S.-M., Twomey, K., Von Holzen, K., Wang, Y., Waxman, S., Werker, J.F., Wermelinger, S., Woolard, A., Yurovsky, D., Zahner, K., Zettersten, M., Soderstrom, M., 2020. Quantifying sources of variability in infancy research using the infant-directed-Speech preference. Adv. Methods Pract. Psychol. Sci. 3, 24–52. https://doi.org/10.1177/2515245919900809.

Fukui, Y., Ajichi, Y., Okada, E., 2003. Monte Carlo prediction of near-infrared light propagation in realistic adult and neonatal head models. Appl. Opt. 42, 2881–2887.

Gemignani, J., Middell, E., Barbour, R.L., Graber, H.L., Blankertz, B., 2018. Improving the analysis of near-infrared spectroscopy data with multivariate classification of hemodynamic patterns: a theoretical formulation and validation. J. Neural Eng. 15, 045001 https://doi.org/10.1088/1741-2552/aabb7c.

Gervain, J., Mehler, J., Werker, J.F., Nelson, C.A., Csibra, G., Lloyd-Fox, S., Shukla, M., Aslin, R.N., 2011. Near-infrared spectroscopy: a report from the McDonnell infant methodology consortium. Dev. Cogn. Neurosci. 1, 22–46. https://doi.org/10.1016/j.dcn.2010.07.004.

Gervain, J., Berent, I., Werker, J.F., 2012. Binding at birth: the newborn brain detects identity relations and sequential position in speech. J. Cogn. Neurosci. 24, 564–574. https://doi.org/10.1162/jocn_a_00157.

Hakuno, Y., Hata, M., Naoi, N., Hoshino, E., Minagawa, Y., 2020. Interactive live fNIRS reveals engagement of the temporoparietal junction in response to social contingency in infants. Neuroimage 218, 116901. https://doi.org/10.1016/j.neuroimage.2020.116901.

Hocke, L.M., Oni, I.K., Duszynski, C.C., Corrigan, A.V., Frederick, B., de, B., Dunn, J.F., 2018. Automated processing of fNIRS data-A visual guide to the pitfalls and consequences. Algorithms 11, 1–25. https://doi.org/10.3390/a11050067.

Hu, X.-S., Arredondo, M.M., Gomba, M., Confer, N., DaSilva, A.F., Johnson, T.D., Shalinsky, M., Kovelman, I., 2015. Comparison of motion correction techniques applied to functional near-infrared spectroscopy data from children. J. Biomed. Opt. 20, 126003 https://doi.org/10.1117/1.jbo.20.12.126003.

Huppert, T.J., 2016. Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. Neurophotonics 3, 010401. https://doi.org/10.1117/1.NPh.3.1.010401.

IBM Corporation, 2017. IBM SPSS Statistics for Windows, Version 25.0.

Issard, C., Gervain, J., 2018. Variability of the hemodynamic response in infants: influence of experimental design and stimulus complexity. Dev. Cogn. Neurosci. 33, 182–193. https://doi.org/10.1016/j.dcn.2018.01.009.

Jackson, E.S., Wijeakumar, S., Beal, D.S., Brown, B., Zebrowski, P., Spencer, J.P., 2019. A fNIRS investigation of speech planning and execution in adults who stutter. Neuroscience 406, 73–85. https://doi.org/10.1016/j.neuroscience.2019.02.032.

Karen, T., Kleiser, S., Ostojic, D., Isler, H., Guglielmini, S., Bassler, D., Wolf, M., Scholkmann, F., 2019. Cerebral hemodynamic responses in preterm-born neonates to visual stimulation: classification according to subgroups and analysis of frontotemporal–occipital functional connectivity. Neurophotonics 6, 1. https://doi.org/10.1117/1.nph.6.4.045005.

Khan, R.A., Naseer, N., Saleem, S., Qureshi, N.K., Noori, F.M., Khan, M.J., 2020. Cortical tasks-based optimal filter selection: an fNIRS study. J. Healthc. Eng. 2020 https://doi.org/10.1155/2020/9152369.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, Š., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E.M., Hasselman, F., Hicks, J.A., Hovermale, J.F., Hunt, S.J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J.A., Barry Kappes, H., Krueger, L.E., Kurtz, J., Levitan, C.A., Mallett, R.K., Morris, W.L., Nelson, A.J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A.M., Schmidt, K., Skorinko, J.L., Smith, R., Steiner, T.G., Storbeck, J., Van Swol, L.M., Thompson, D., van 't Veer, A.E., Ann Vaughn, L., Vranka, M., Wichman, A.L., Woodzicka, J.A., Nosek, B.A., 2014. Investigating variation in replicability: a "ManyLabs" replication project. Soc. Psychol. (Gott). 45, 142–152. https://doi.org/10.1027/1864-9335/a000178.

Lloyd-Fox, S., Blasi, A., Volein, A., Everdell, N., Elwell, C.E., Johnson, M.H., 2009. Social perception in infancy: a near infrared spectroscopy study. Child Dev. 80, 986–999. https://doi.org/10.1111/j.1467-8624.2009.01312.x.

Lloyd-Fox, S., Blasi, A., Elwell, C.E., Charman, T., Murphy, D., Johnson, M.H., 2013. Reduced neural sensitivity to social stimuli in infants at risk for autism. Proc. R. Soc. B Biol. Sci. 280 https://doi.org/10.1098/rspb.2012.3026.

Lloyd-Fox, S., Begus, K., Halliday, D., Pirazzoli, L., Blasi, A., Papademetriou, M., Darboe, M.K., Prentice, A.M., Johnson, M.H., Moore, S.E., Elwell, C.E., 2017. Cortical specialisation to social stimuli from the first days to the second year of life: a rural Gambian cohort. Dev. Cogn. Neurosci. 25, 92–104. https://doi.org/10.1016/j.dcn.2016.11.005.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.

McDonald, N.M., Perdue, K.L., Eilbott, J., Loyal, J., Shic, F., Pelphrey, K.A., 2019. Infant brain responses to social sounds: a longitudinal functional near-infrared spectroscopy study. Dev. Cogn. Neurosci. 36, 100638 https://doi.org/10.1016/j.dcn.2019.100638.

Mercure, E., Evans, S., Pirazzoli, L., Goldberg, L., Bowden-Howl, H., Coulson-Thaker, K., Beedie, I., Lloyd-Fox, S., Johnson, M.H., MacSweeney, M., 2020. Language experience impacts brain activation for spoken and signed language in infancy: insights from unimodal and bimodal bilinguals. Neurobiol. Lang. 1, 9–32. https://doi.org/10.1162/nol_a_00001.

Miguel, H.O., Lisboa, I.C., Gonçalves, O.F., Sampaio, A., 2019. Brain mechanisms for processing discriminative and affective touch in 7-month-old infants. Dev. Cogn. Neurosci. 35, 20–27. https://doi.org/10.1016/j.dcn.2017.10.008.

Miguel, H.O., Gonçalves, Ó.F., Sampaio, A., 2020. Behavioral response to tactile stimuli relates to brain response to affective touch in 12-month-old infants. Dev. Psychobiol. 62, 107–115. https://doi.org/10.1002/dev.21891.

Minagawa, Y., Hakuno, Y., Kobayashi, A., Naoi, N., Kojima, S., 2017. Infant word segmentation recruits the cerebral network of phonological short-term memory. Brain Lang. 170, 39–49. https://doi.org/10.1016/j.bandl.2017.03.005.

Minagawa-Kawai, Y., Mori, K., Hebden, J.C., Dupoux, E., 2008. Optical imaging of infants' neurocognitive development: recent advances and perspectives. Dev. Neurobiol. 68, 712–728. https://doi.org/10.1002/dneu.20618.

Molavi, B., Dumont, G.A., 2012. Wavelet-based motion artifact removal for functional near-infrared spectroscopy. Physiol. Meas. 33, 259–270. https://doi.org/10.1088/0967-3334/33/2/259.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349. https://doi.org/10.1126/science.aac4716 aac4716.

Peña, M., Maki, A., Kovačić, D., Dehaene-Lambertz, G., Koizumit, H., Bouquet, F., Mehler, J., 2003. Sounds and silence: An optical topography study of language recognition at birth. Proc. Natl. Acad. Sci. U. S. A. 100, 11702–11705. https://doi.org/10.1073/pnas.1934290100.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical Parametric Mapping: the Analysis of Functional Brain Images. Academic press.

Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., Tachtsidis, I., 2019. Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. Front. Hum. Neurosci. 12, 1–21. https://doi.org/10.3389/fnhum.2018.00505.

Porto, J.A., Bick, J., Perdue, K.L., Richards, J.E., Nunes, M.L., Nelson, C.A., 2020. The influence of maternal anxiety and depression symptoms on fNIRS brain responses to emotional faces in 5- and 7-month-old infants. Infant Behav. Dev. 59, 101447 https://doi.org/10.1016/j.infbeh.2020.101447.

Roche-Labarbe, N., Fenoglio, A., Aggarwal, A., Dehaes, M., Carp, S.A., Franceschini, M. A., Grant, P.E., 2012. Near-infrared spectroscopy assessment of cerebral oxygen metabolism in the developing premature brain. J. Cereb. Blood Flow Metab. 32, 481–488. https://doi.org/10.1038/jcbfm.2011.145.

Santosa, H., Zhai, X., Fishburn, F., Huppert, T., 2018. The NIRS brain AnalyzIR toolbox. Algorithms 11, 73. https://doi.org/10.3390/a11050073.

Scholkmann, F., Spichtig, S., Muehlemann, T., Wolf, M., 2010. How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation. Physiol. Meas. 31, 649–662. https://doi.org/10.1088/0967-3334/31/5/004.

Scholkmann, F., Kleiser, S., Metz, A.J., Zimmermann, R., Mata Pavia, J., Wolf, U., Wolf, M., 2014. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. Neuroimage 85, 6–27. https://doi.org/10.1016/j.neuroimage.2013.05.004.

Taga, G., Watanabe, H., Homae, F., 2018. Developmental changes in cortical sensory processing during wakefulness and sleep. Neuroimage 178, 519–530. https://doi.org/10.1016/j.neuroimage.2018.05.075.

Ujiie, Y., Kanazawa, S., Yamaguchi, M.K., 2020. The other-race-Effect on audiovisual speech integration in infants: a NIRS study. Front. Psychol. 11, 1–12. https://doi.org/10.3389/fpsyg.2020.00971.

Van Der Kant, A., Männel, C., Paul, M., Friederici, A.D., Höhle, B., Wartenburger, I., 2020. Linguistic and non-linguistic non-adjacent dependency learning in early development. Dev. Cogn. Neurosci. 45, 100819 https://doi.org/10.1016/j.dcn.2020.100819.

Wilcox, T., Bortfeld, H., Woods, R., Wruck, E., Boas, D.A., 2005. Using near-infrared spectroscopy to assess neural activation during object processing in infants. J. Biomed. Opt. 10, 011010 https://doi.org/10.1117/1.1852551.

Wilcox, T., Haslup, J.A., Boas, D.A., 2010. Dissociation of processing of featural and spatiotemporal information in the infant cortex. Neuroimage 53, 1256–1263. https://doi.org/10.1016/j.neuroimage.2010.06.064.

Xu, Y., Graber, H.L., Barbour, R.L., 2014. nirsLAB: a computing environment for fNIRS neuroimaging data analysis. Biomed. Opt. Express 2014. https://doi.org/10.1364/BIOMED.2014.BM3A.1. BM3A.1.

Yücel, M.A., Selb, J., Cooper, R.J., Boas, D.A., 2014. Targeted principle component analysis: a new motion artifact correction approach for near-infrared spectroscopy. J. Innov. Opt. Health Sci. 7, 1–8. https://doi.org/10.1142/S1793545813500661.

Zimmermann, B.B., Roche-Labarbe, N., Surova, A., Boas, D.A., Wolf, M., Grant, P.E., Franceschini, M.A., 2012. The confounding effect of systemic physiology on the hemodynamic response in newborns. Adv. Exp. Med. Biol. 737, 103–109. https://doi.org/10.1007/978-1-4614-1566-4_16.