Research paper

Use of eye tracking improves the detection of evoked responses to complex visual stimuli during EEG in infants

Eero Ahtola [a,b,*], Susanna Stjerna [a], Nathan Stevenson [a], Sampsa Vanhatalo [a]

[a] Department of Children's Clinical Neurophysiology, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
[b] Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland

ARTICLE INFO

ABSTRACT

Objective: To improve the reliability of detecting EEG responses evoked by complex visual stimuli to the level required for clinical use by integrating an eye tracker to the EEG setup and optimizing the analysis protocol.
Methods: Infants were presented with continuous orientation reversal (OR), global form (GF), and global motion (GM) stimuli. Eye tracking was used to control stimulus presentation and exclude epochs with disoriented gaze. The spectral responses were estimated from 13 postcentral EEG channels using a circular variant of Hotelling's T^2 test statistic.
Results: Among 39 healthy infants, statistically significant ($p < 0.01$) responses to OR/GF/GM stimuli were found from 92%/100%/95% recordings, respectively. The specificity test of the detection algorithm, using non-stimulated baseline EEG, did not yield any false-positive findings. Taken together, this yields 15% improvement on average in the detection performance compared to that in the current literature.
Conclusions: Changes to the test protocol and incorporation of the eye tracking information improves the detection of responses to complex visual stimuli in infants.
Significance: This work presents a test protocol suitable for use in a clinical environment at a level of reliability that allows individual diagnostics.

1. Introduction

Progress in neonatal care has improved survival and reduced the incidence of major neurological sequalae (Saigal and Doyle, 2008; Sellier et al., 2010); however, the mild effects of early adversities on later neurocognitive development remain a major challenge in developmental neurology (Johnson and Marlow, 2016). Early detection of developmental deviances is crucial for guiding individualized treatment and rehabilitation. Several standardized neurological assessment scales and neuroimaging and neurophysiological methods are used by clinicians for the early detection of compromised function (Merchant and Azzopardi, 2015).

Neurocognitive performance and complex visual processing in childhood are commonly affected by neonataladversities

(Aarnoudse-Moens et al., 2009; Braddick and Atkinson, 2011; Johnson et al., 2016), and several lines of evidence suggest that assessment of the visual system will provide a proxy for assessing wider neurocognitive development or its risks. Detection of the change in orientation requires cortical neuronal activity (Braddick and Atkinson, 2011), and therefore, it can be used for studying the integrity of the visual pathway from the retina to the primary visual cortex, which is important in more severe neonatal adversities. Aspects of more complex visual processing may offer proxy to assess readiness for visual processing-related and reliant neurocognitive functions across multiple cortical areas beyond the primary visual cortex, e.g., the clustering of cortical areas referred to as dorsal and ventral streams (Braddick et al., 2000, 2003; Braddick and Atkinson, 2011; Milner, 2012).

Dorsal stream processing, including temporoparietal areas, is a prerequisite for global motion (GM) perception and closely linked to visuomotor and visuospatial skills and attentional processes (Braddick et al., 2000, 2003; Wattam-Bell et al., 2010; Milner, 2012). In addition to developmental disorders and dyslexia (Grinter et al., 2010; Braddick and Atkinson, 2011; Milner, 2012; Robertson et al., 2014), deficits in GM perception have been recently associated with mathematical learning difficulties in

childhood (Braddick et al., 2016). In contrast, global form (GF) processing is believed to, at least partly, rely on distinct ventral stream functioning (Braddick et al., 2000) required for object recognition, including faces (Braddick and Atkinson, 2011). It has also been suggested that ventral stream processing activates brain mechanisms for perceptual processes that serve planning, recognition, and memory (Milner, 2012), and difficulties in GF processing have been associated, for instance, with dyspraxia (O'Brien et al., 2002; Grinter et al., 2010). Finally, in addition to the efficient functioning of both the streams, interaction between them is required for developing skills of flexible attention regulation and neurocognitive processing (Grinter et al., 2010; Milner, 2012; Binkofski and Buxbaum, 2013; Cloutman, 2013).

Studying complex visual processing typically relies on behavioral responses that may be difficult to achieve in preverbal infants, especially in neurologically compromised subjects. Intriguingly, recording cortical responses to complex visual stimuli may elucidate higher level cortical processing as early as few months of age (Braddick et al., 1986). This EEG response to visual stimulus (ERVS) method was introduced by Braddick et al. (1986, 2000) and Wattam-Bell et al. (2010) and is based on analyzing cortical steady-state responses to a periodic visual stimulus using orientation reversal (OR), GF, and GM paradigms. These paradigms are considered to reflect cortical visual processing at the level of the V1 area (OR) and the ventral (GF) and dorsal (GM) streams (Braddick et al., 2000, 2005; Wattam-Bell et al., 2010). These visual evoked responses have been subsequently studied in several infant groups who were at risk of developmental compromise, including very or extremely preterm babies (Atkinson et al., 2002) and infants with perinatal brain injuries (Mercuri et al., 1996) or other neurodevelopmental disorders (Braddick and Atkinson, 2011; Lee et al., 2012).

While the test paradigm has been used in several studies by the original authors, its wider spread has been slow, mainly because of the methodological issues that make its clinical implementation a challenge. Most importantly, there have been inadequate means to control stimulation according to subject's attention, the analysis scripts have not been openly available, and the stimulation system has relied on experienced and manual operation. Prior studies have also shown that the existing method of implementation may be unable to detect responses even in healthy infants. This suggests low sensitivity, which significantly challenges the ERVS method's use in clinical diagnostics.

In the present work, we aimed to improve the ERVS method to enable reliable studies as part of the clinical routine in our pediatric neurophysiological unit. We believe that several problems with the ERVS paradigm may be solved. First, the reported low sensitivity could be due to poor coordination between the patient's gaze and stimulus presentation, leading to analysis of signals when the infant is not observing the screen. This was addressed by implementing continuous gaze tracking to guide the stimulus presentation and analysis. Second, factors related to low electrode count (e.g., only one occipital channel as used by Atkinson et al., 2002; Braddick et al., 2005; Lee et al., 2012) or suboptimal computational paradigm could lead to false-negative results. In the present study, we thoroughly investigated the experimental variables to maximize the response rate in healthy infants.

2. Materials and Methods

2.1. Participants

We recruited a cohort (N = 39; 17 females) of normally developing infants (according to medical history and parent interview) that were born at full term (>37 weeks of gestation) at the Helsinki University Hospital. The results were recorded at mean age of 3.4 months (range 3.0–4.0 months). Two infants were re-tested at the age of 5.7 (subject 23) and 5.8 (subject 25) months because of technical difficulties in the first test session.

2.2. Devices and recordings

2.2.1. Eye tracking

The eye tracker system used in our study was Tobii T120 (Tobii Technology AB, Stockholm, Sweden) and is equipped with an integrated 17'' thin-film transistor display (refresh rate: 60 Hz, response time: 4 ms). The system samples gaze data at 120 Hz, operates at 50–80 cm distance from the eyes, and can follow head movements within a window of 30×22 cm (at 70 cm from the screen).

The Tobii eye tracker system is based on pupil center corneal reflection, i.e., near-infrared illumination and its reflections from the cornea relative to the center of the pupil. The light reflections are captured by two cameras, and a general 3D model of the eye and the angles, distances, and other geometrical features of the reflections are used to calculate the positions of the eyes and the direction of gaze. This can then be projected onto the display to identify the pixel coordinates where the gaze is fixated.

2.2.2. EEG recording and re-referencing

The visual responses were collected by a multichannel EEG recording. We used the routine NicoletOne V44 EEG recorder (Cardinal Healthcare/Natus, U.S.A.) and ANT WaveGuard EEG caps (ANT-Neuro, Germany) with 20 Ag/AgCl electrodes positioned according to the international 10–20 electrode system. Because the analysis was focused on posterior–occipital regions, we also included Oz. We used 5 kΩ as the target impedance level at the electrode–skin interface before data acquisition. The sampling rate was 500 Hz, and Cz was used as the recording reference. After recording, the data were exported as an EDF file for further processing using custom-scripted MATLAB routines.

We compared multiple different offline montages. Because the evoked responses were expected in the postcentral regions, we performed the primary analysis using a regional frontal average reference computed from the five frontal EEG channels (F3, F4, F7, F8, and Fz). This choice was made to reduce the sensitivity to single electrode noise, which may readily confound monopolar recordings (as shown in Supplementary Fig. A.1). For instance, frontal muscle activity is unavoidably present in awake infants, but its lack of spatial coherence (Freeman et al., 2003) allows its easy removal by averaging. In addition, we also performed analyses using (i) simple Fz reference, (ii) the original Cz-referenced recording montage (comparable to Braddick et al., 2005), and (iii) the commonly used grand average montage (Wattam-Bell et al., 2010) including all channels (except Fp1 and Fp2) in the reference. Channels with a standard deviation of >300 μV over the whole recording were excluded from both average references.

2.2.3. Synchronization between stimuli, EEG, and eye tracking

The start and end times of the video files were time-stamped to the gaze data files to allow synchronization between stimuli and eye tracking. Synchronization between stimuli and EEG was achieved using a custom battery-powered light sensor (BPW34 photodiode; Vishay Semiconductors, U.S.A) attached on the corner of the stimulation screen. We incorporated black and white squares in this corner (visible in Fig. 1A and B) to provide a frame-level timing of the visual stimuli. The light sensor transformed the binary patterns into pulses for a bipolar input of the EEG amplifier. In later studies, this custom-made light sensor was replaced by a commercially available, photodiode-based StimTracker (Cedrus, CA, U.S.A) that transfers trigger information

directly to the EEG amplifier as transistor–transistor logic-compatible digital pulses.

2.3. Stimuli

2.3.1. Orientation reversal

The OR stimulus consists of high-contrast sinusoidal wave gratings with a spatial frequency of 0.45 cycles/degree at an average viewing distance of 60 cm. The grating pattern switches its orientation back and forth between oblique angles of 45° and 135° at a frequency of 4 Hz (Fig. 1B top; Braddick et al., 1986, 2005).

The ORs are accompanied by random phase shifts in the gratings recurring at 24 Hz. The purpose of these "jitters" is to control non-orientation-specific responses to local contrast change. Because of the random phase shifts, local luminance variations in the pattern are statistically similar in every jitter and OR and do not contribute to the component at the stimulus frequency (Braddick et al., 2005).

2.3.2. Global form

The GF stimulus consists of a set of short, white arc segments on a black background (N = 2000, length = 1.3°). The arrangement of the arcs switches between global coherent and non-coherent states at a frequency of 2 Hz. In the coherent state, the arcs are concentrically organized to create a global circular pattern, while in the non-coherent state, the arcs are randomly oriented without any perceivable global structure (Fig. 1B middle; Wattam-Bell et al., 2010).

2.3.3. Global motion

The GM stimulus consists of a set of white dots (N = 2000, diameter = 0.3°) that move on a black background. Similar to GF, the stimulus alternates between states of global coherence and non-coherence at a frequency of 2 Hz. In the coherent state, the dots move along short tangential trajectories creating the perception of a global rotational movement, and in the non-coherent state, they move independently in random directions (Fig. 1B bottom; Wattam-Bell et al., 2010).
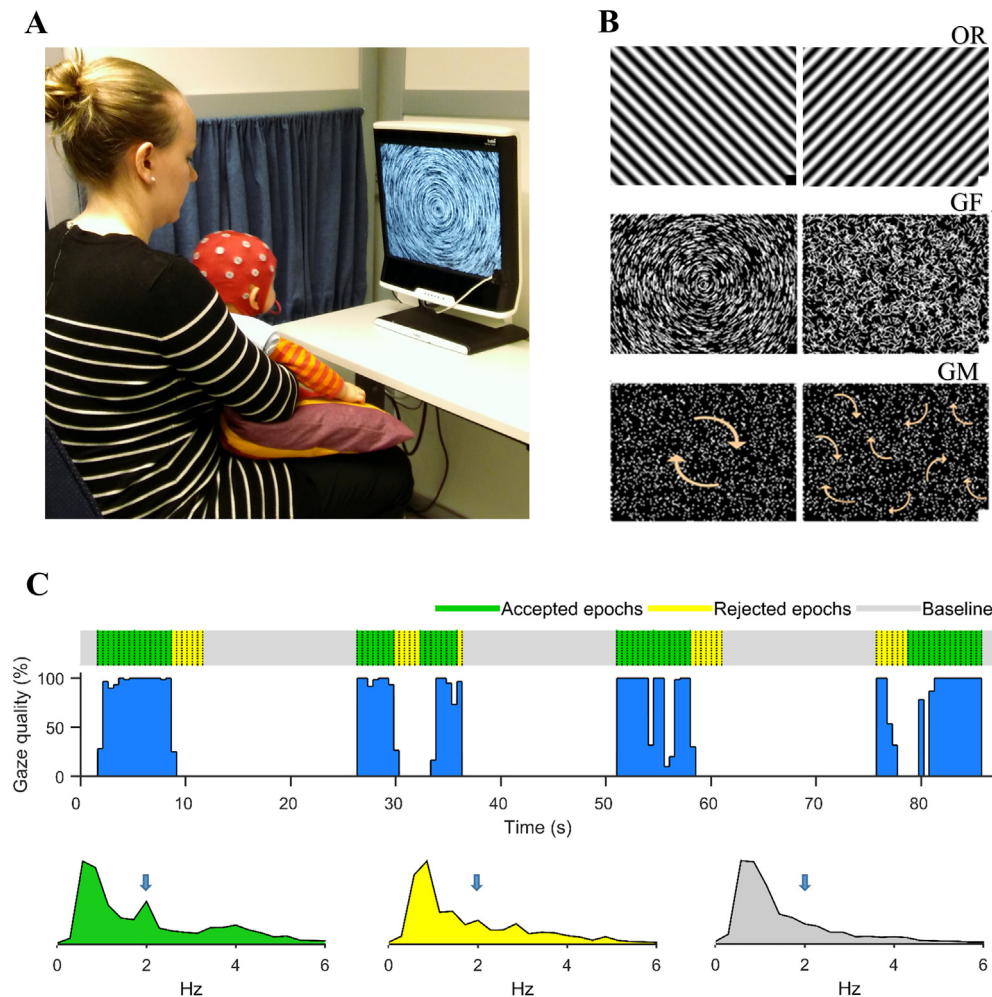
Fig. 1. Overview of the paradigm. **A** Infant sits in the lap of the caregiver and is shown visual stimuli continuously on the display of the eye tracker that is synchronized to a parallel EEG recording through a light sensor attached on the corner of the screen. The experimenter sits behind a light wall and controls the progression of the test according to the observations through a one-way transparent window and an interactive feedback mechanism of the presentation software. **B** In the orientation reversal (OR, top) stimulus, sine grating switches its orientation continuously between 45° and 135° angles. In the global form (GF, middle) and motion (GM, bottom) stimulation, coherent and non-coherent phases are presented alternately. The stimulus consists either of short white arc segments (GF) or small white dots (GM). **C** Above: The blue bars depict the evolution of gaze fixation quality during four successive 10-s-long GM stimulation sequences, each consisting of 20 reversal cycles (black dotted lines). The occasional gaze fixation quality drops indicate the participant's temporary disorientation from the stimulus. The epoch selection algorithm identifies segments with good (green) and poor (yellow) gaze fixation for further analysis (fixed gaze threshold of 45% used here). Below: Power spectrum of good fixation epochs (green) reveals a distinct peak at a stimulus frequency of 2 Hz. The spectral peak is substantially decreased in rejected epochs with poor fixation (yellow) and completely missing from the spectrum calculated from the baseline epochs measured in between the stimulus sequences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3.4. Generation of stimuli as video files

All stimuli were in digital video format, first generated frame-by-frame in MATLAB using the routines included in Psychophysics toolbox (version 3; Brainard, 1997). The frames were then encoded into MPEG-4 video files of 10, 20, and 30 s lengths. The frame rates of the video clips were 24 Hz for OR and 60 Hz for GF and GM stimuli. The resolution was 1024 × 768 pixels for OR and 800 × 600 for GF and GM because of their higher frame rate. Low-resolution versions of the videos are shown in the Supplementary Material (Appendix B.1). The video files with original resolution are available on request from the author.

2.4. Test session

During the recording session (see Fig. 1A), the infant was connected to an EEG recorder and placed in the sitting position in a baby carrier attached on their parent's chest in front of the eye tracker display. The ambient light was kept dim to reduce the interference of diffuse light reflections with the eye tracker, and the infant-parent dyad was separated from the experimenter with a light wall to improve the baby's concentration. The parents were informed not to look at the screen to avoid confusing the eye tracker.

Sessions always began with a routine calibration procedure (Tobii Studio: version 2.2.8). Then, for each of the three visual stimuli, we aimed to record at least 100 s of EEG from periods where gaze quality was of sufficiently high quality (accumulation monitored by the eye tracker). In some infants where this was not achievable because of poor cooperation, we prioritized the GF and GM stimuli over OR.

2.4.1. Fixation-based stimulus presentation

The stimuli were presented using the E-Prime software (version 2.0.8.22, Psychology Software Tools, PA, U.S.A) and E-Prime Extensions for Tobii (version 2.0.1.5) interfacing with the eye tracker hardware.

The presentation of visual stimuli was controlled by the infant's gaze. We used an attractor, a simple video animation presented in the center of the screen, and the playback of stimulation video started only after the eye tracker had detected a sufficiently stable (600-ms-long segment of continuous fixation on the target) fixation on it. The operator could also see a real-time presentation of the infant's eye positions, which could be used to manually launch the stimulus presentation with the pressing of a key.

2.5. Response analysis

The rationale of our data analysis was that continuous visual stimulation elicits a steady-state response in the visual cortex that may be distinguished as a significant increase in spectral energy at the stimulus frequency compared to EEG where no stimulus is applied (Braddick et al., 2005). Notably, this effect is much milder than the typical flash-evoked photic driving; moreover, the recording technician can recognize the response by online visual inspection of the EEG signal, although only occasionally (example in Supplementary Fig. A.1). Detection of the spectral peak may be accentuated by incorporating an eye tracker-based segmentation (Fig. 1C). The individual response rate was statistically examined for each stimulus type. The null hypothesis was that the spectral values at the stimulus frequency, across several epochs of the recorded signal, were consistent with random fluctuations (one-sample test) or indifferent from components of a separate baseline sample taken from the non-stimulation phases of the same recording (two-sample test). Conversely, the alternate hypothesis was that this difference was significant and a spectral component

related to the evoked response was present in the signal within a given confidence level.

In addition, we performed spatial analysis of the detected responses using the multichannel EEG data from the electrodes positioned across the scalp to (i) ascertain that the responses occur in regions known to involve visual processing, (ii) improve the response detection algorithm by focusing on these areas, and (iii) guide appropriate EEG montage.

Fig. 2 presents a flow chart of the analysis chain that was used in the work starting from the raw EEG and simultaneous eye tracker data to the determination of the response presence. The essential signal processing steps are shared as a downloadable MATLAB package in the Supplementary Information C.1.
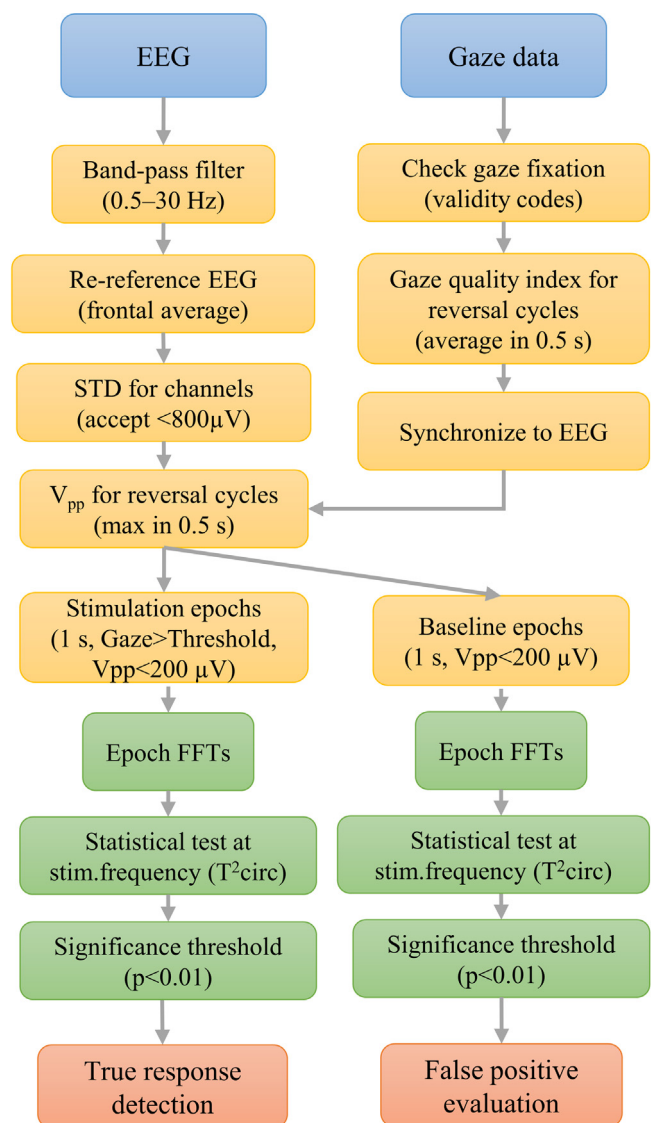
Fig. 2. Flow chart of the method presented in this work. The chart presents the signal processing pipelines that were used in the present work, starting from the raw EEG and eye tracker gaze data to the determination of the response presence using a statistical test. Boxes show all the major processing steps, and their specific parameters are given in parenthesis. The signal processing steps on green-colored blocks can be found in the Supplementary Material of this paper as MATLAB scripts along with brief example dataset of the recordings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.5.1. EEG pre-processing and epoch segmentation

After re-referencing, all EEG channels were processed independently. We first filtered the EEG signals with an eighth-order Butterworth bandpass filter (passband 0.5–30 Hz) with forward and reverse implementation (Braddick et al., 2005; Wattam-Bell et al., 2010). Next, channels with excessively poor signal quality (e.g., loose electrode contact) were excluded from further processing if their standard deviation exceeded 800 μV.

The epoch segmentation was initiated by assessing the EEG of all 0.5-s-long stimulus reversal cycles on the basis of the recorded trigger data. In a simple artefact rejection procedure, the segments were rejected if their peak-to-peak voltage exceeded 200 μV (Braddick et al., 2005; Wattam-Bell et al., 2010).

The accepted EEG segments were then merged to form longer epochs. This was performed using a sliding window (0.5 s step) that calculated the proportion of time when the gaze was directed on the screen according to the eye tracker. This gaze quality index ranged from 0% to 100% and was applied to a fixed threshold (Fig. 1C). The tracked gaze coordinates (x and y) were not used in response analyses.

We mainly used epoch length of 1 s with 45% gaze quality threshold, while these parameters were systematically assessed over a wide range of values. The spectral estimate of the EEG was computed using Fourier transform (FFT), which, with 1 s epochs, provides a frequency resolution of 1 Hz.

Finally, we selected baseline epochs from periods without visual stimulation (usually before, after, or in between the stimulus sequences), and the rejection criteria (>200 μV peak-to-peak voltage) was applied for epochs with stimulation. This epoch set acted as an unstimulated reference in the method evaluation and in the two-sample statistical tests.

2.5.2. Computational response detection methods

As with the previous studies using OR, GF, and GM stimuli (Braddick et al., 2005; Wattam-Bell et al., 2010), we used a circular variant of T^2 test statistic by Hotelling (1931) called T^2circ (Victor and Mast, 1991). This statistic measures the signal-to-noise ratio of periodic events in the recorded EEG. Higher T^2circ values indicate less variability between the epochs and thus stronger correlation between the stimulation and the recording.

T^2circ is a bivariate statistic that utilizes the real and imaginary components of the spectral representation (through the FFT) of the EEG signal at the stimulus frequency, which is assumed to be independent. In the one-sample version of T^2circ, the null hypothesis of the test is that no response signal is present, i.e., the spectral properties of the EEG epochs are indifferent to random scatter about zero (noise). Then, T^2circ may be calculated from N epochs of EEG using the following equation:

$$T^2_{circ} = N(N-1) \frac{|\langle z \rangle_{est}|^2}{\sum_{j=1}^{N} |z_j - \langle z \rangle_{est}|^2}, \tag{1}$$

where z_j is the independent and complex-valued spectral values at a given stimulus frequency (2 or 4 Hz) and $\langle z \rangle_{est}$ is their empirical mean across epochs.

When defined this way, T^2circ is statistically distributed according to a F-distribution ($F_{[2,2N-2]}$), which allows the determination of a p-value. Moreover, it induces a dependence on the number of analyzed epochs. When the size of a dataset consisting of N statistically uniform epochs is increased, the resulting T^2circ values monotonically increase as a function of N (see T^2circ tests with simulated EEG data in the Supplementary Fig. A.2). This is expected in hypothesis testing as smaller differences can be deemed significant when the sample size increases; however, this is an issue as we have variable epoch numbers per infant.

In addition to the conventional T^2circ, several other statistical tests were evaluated: one-sample univariate t (Victor and Mast, 1991), one-sample multivariate Hotelling's T^2 (Hotelling, 1931), two-sample univariate Mann-Whitney U (Mann and Whitney, 1946), two-sample univariate t, two-sample multivariate Hotelling's T^2 (Hotelling, 1931), and two-sample multivariate T^2circ (Victor and Mast, 1991). In the two-sample tests, the reference sample consisted of the baseline epochs. The multivariate statistics exploit both the real and imaginary spectral components at the stimulus frequency, whereas univariate tests only use the magnitude of these components.

2.5.3. Detection from multichannel data

Visual responses are conventionally studied using only one recording electrode placed over the occipital area (e.g., at Oz; Odom et al., 2010; Norcia et al., 2015), and the reference electrode is typically placed at the vertex or midfrontally. This fixed placement is very sensitive to inter-subject variability in response topography (Picton et al., 2003). Our prior work on spatial properties in infant EEG has suggested that the scalp-recorded EEG activity may significantly vary within few centimeters (Odabaee et al., 2013, 2014). Hence, using multichannel data is likely to improve response detection; however, it also leads to the need for the appropriate correction of multiple statistical comparisons to control family-wise error rate (i.e., positive findings by chance).

We adopted the method in the reference works: the false discovery rate (FDR) correction by Benjamini and Hochberg (1995) that adjusts the selected confidence threshold of the statistical test according to the number of parallel comparisons—in this case, the number of channels analyzed. To avoid performing an overly conservative correction, we decided to limit the number of the channels analyzed by focusing on the postcentral area alone for response detection (specifically T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, Oz, and O2 were included).

2.5.4. Evaluation of the response detection performance

One of the aims of this study was to investigate the various technical factors that could affect the response detection results. In the analysis workflow presented above, variables such as the number of EEG channels, the re-referencing method, the applied statistical test, p-value criterion, epoch length, and the gaze quality threshold of the epoch formation must all be chosen carefully to provide optimal results.

This optimality means a combination of the experimental variables that maximizes the sensitivity and specificity of the classifier (response detector). As there was no "ground truth" information available about the findings in our cohort, we had to assume that all babies elicited a response during stimulus intervals and no response during the non-stimulus. Thus, sensitivity was defined as the percentage of infants with a response at the stimulus frequency during stimulus (true-positive detection rate; TPR) and specificity as the percentage of infants with no response at the stimulus frequency during non-stimulus (true-negative detection rate; TNR).

We used the area under receiver operating characteristic curve (AUC) to combine the measures of sensitivity and specificity. Single-point AUC is defined as

$$AUC = \frac{TPR - FPR + 100\%}{2}, \tag{2}$$

where FPR is the false-positive rate, which equals 100% − TNR. Infants were included if at least 60 s of valid EEG data were recorded during stimulus.

2.5.5. Response topography

Spatial distributions of the visual responses were examined using all the 20 electrodes. Each channel was analyzed individually. After pooling studies with significant responses, the group-level topographic findings were presented in two ways, separately for each stimulus type: (i) by averaging the calculated T^2circ values for each cohort, and (ii) by calculating the channel-specific response rates that show how many of recordings yielded a response for each channel.

3. Results

3.1. Response rates after different detection methods

We first analyzed the EEG using the settings originally proposed by Braddick et al. (2005) (T^2circ statistics; $p < 0.05$; 1 s epochs; only Oz) ignoring the gaze data from the eye tracker, which resulted in AUCs of 92%, 94%, and 78% for OR, GF, and GM, respectively. Incorporating multichannel EEG data (13 postcentral electrodes) and corrections for multiple comparisons improved the AUCs to 94%, 97%, and 92%, respectively. By reducing the statistical detection criterion (p-value), we could alter the ratio of false positives to sensitivity. A detection threshold of $p < 0.01$ resulted in AUCs of 96%, 99%, and 92% for OR, GF, and GM, respectively, with a FPR of 0% for all stimulus types.

The use of longer epoch length (Supplementary Fig. A.3) or alternative one-sample statistical tests did not significantly improve the AUCs. Hotelling's T^2 (AUCs 94–97%) provided almost identical figures for T^2circ, and even conventional Student's *t*-test showed reasonable performance (AUCs 91–96%). Two-sample tests incorporating a set of baseline EEG epochs did not increase the response rate compared to one-sample tests (best was Hotelling's T^2: TPRs 79–92%). The comprehensive detection results acquired using the other statistical tests described in the section "computational response detection methods" are presented as a table in Supplementary Information C.2.

3.2. Incorporating eye tracking into analysis

Our initial analyses suggested that epoch rejection based on the gaze quality index can result in larger T^2circ values, indicating an improvement in detection algorithm efficiency (see Fig. 3 and Supplementary Fig. A.4). To maximize the T^2circ per subject, a variable threshold would be required as there was significant variation with subjects and stimulus types (Fig. 3D).

We performed epoch rejection using a variable gaze threshold that was automatically selected to maximize the T^2circ in each case (0–95% in 5% steps, with the additional condition that N > 60). The total epoch quantities before and after this are presented as a table in Supplementary Information C.3. With a detection threshold of $p < 0.01$ applied, this approach resulted in AUCs of 96%, 100%, and 97% for OR, GF, and GM, respectively, no false positives, and an average rejection rate below 1%.

The increase in AUCs with regard to changes in the analysis method is illustrated in Fig. 4, whereas the findings are presented individually in the Supplementary Fig. A.5 together with detection results calculated using the alternative EEG montages.

3.3. Response topography

Topographical analysis of the responses was performed only at group level (Fig. 5). Each stimulus type elicited a distinct response mainly in the posterior and occipital areas. There were also modest differences in the spatial extent of the response magnitudes and response rates. Typically, the responses for the OR and GF stimula-

tions were more prominent and appeared mainly in the occipital channels, while GM responses covered a broader area including electrodes from the horizontal midline (T3-C3-Cz-C4-T4).

Additional results from topography analyses, consisting of surface maps with alternative montages, are presented in Supplementary Fig. A.6. They show how the use of grand average reference resulted in the apparent spread of response to frontal regions, which was avoided with frontal average and Fz-referenced montages.

4. Discussion

Our work shows that spectral responses to complex visual stimuli of OR, GF, and GM can be reliably recorded from the infant cortex at few months of age. While the overall finding is fully compatible with those of earlier ERVS studies (Braddick et al., 2005; Wattam-Bell et al., 2010), we show here that the response detection rates can be substantially improved by further development of the recording setup and by optimizing the signal analysis pipeline.

The initial aim of our work was to implement the ERVS method of Braddick et al. (2005) and Wattam-Bell et al. (2010) with three major enhancements: (i) use of multichannel EEG for all stimulus types, (ii) integration of eye tracking information, and (iii) optimization of the analysis protocol. The outcome was evaluated using a cohort of healthy 3-month-old infants, and the resulting AUCs of 96%, 100%, and 97% for OR, GF, and GM, respectively, exceeded our expectations.

We found several essential and adjustable details in the analysis flow. We saw improvements in response rate from (i) the use of conservative classification criterion and (ii) epoch segmentation based on gaze quality. The use of the criterion $p < 0.01$ eliminated all false-positive findings and preserved true detection rates between 92% and 100%. The introduction of epoch selection using adaptive thresholding of gaze quality permitted the removal of segments with poor gaze quality and predominantly low statistical values, which further improved the detection, thus increasing AUCs by 2% on average. These results should, however, be interpreted in the context of earlier studies where the response rates ($p < 0.05$) in the corresponding age group were approximately 79%, 50%, and 92% for OR, GF, and GM, respectively (Braddick et al., 2005; Wattam-Bell et al., 2010). Compared to them, our enhancements produced an overall increase of 15% in AUCs on average.

The implemented epoch selection approach should not be confused with the artificial improvement of results. It is, indeed, fully analogous to the manual rejection of artefacts in conventional evoked potential studies (e.g., Luck, 2005; Odom et al., 2010; Picton et al., 2000) that result in improved signal-to-noise ratios. Here, we developed an automated way to perform this optimization, which allows more objective and easier data analysis accounting for the observed variability between the recordings.

In addition to the use of gaze quality for epoch selection, we also found that stimulus presentation can be dynamically guided by the eye tracker to ensure that the infant's attention is on the screen while stimulating. Such co-operation is obviously crucial for the assessment of visual responses. Our experimenters also found it highly beneficial to have a real-time feedback of infant's gaze during stimulation so that they could manually control the system as needed. We were unable to directly assess the potential improvement as we did not perform experiments without the use of the eye tracker. However, it certainly helped us to record an adequate amount of EEG from nearly all the participants (>60 epochs in 114/117 sessions) and may also be a reason why our initial analyses exceeded the figures reported in prior studies.
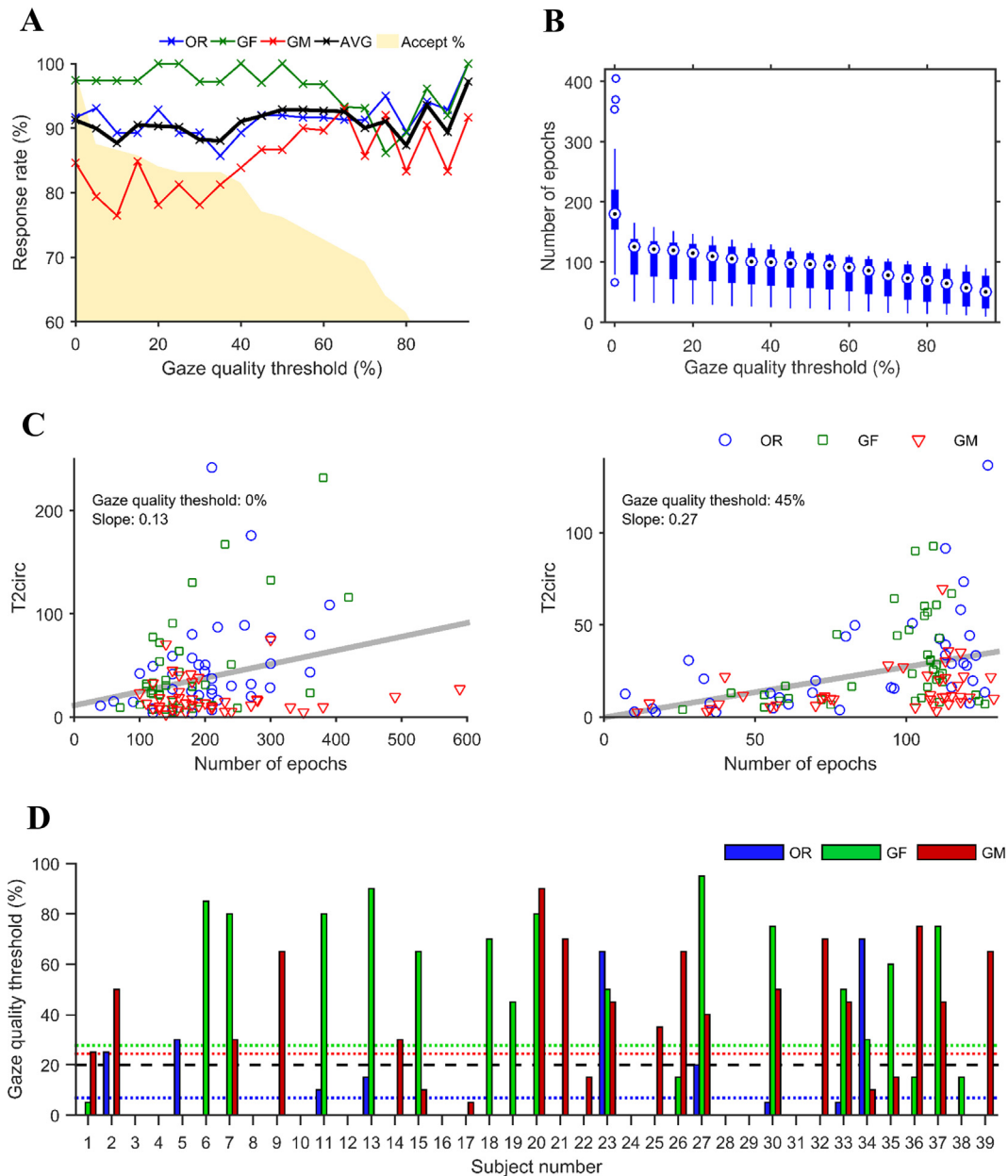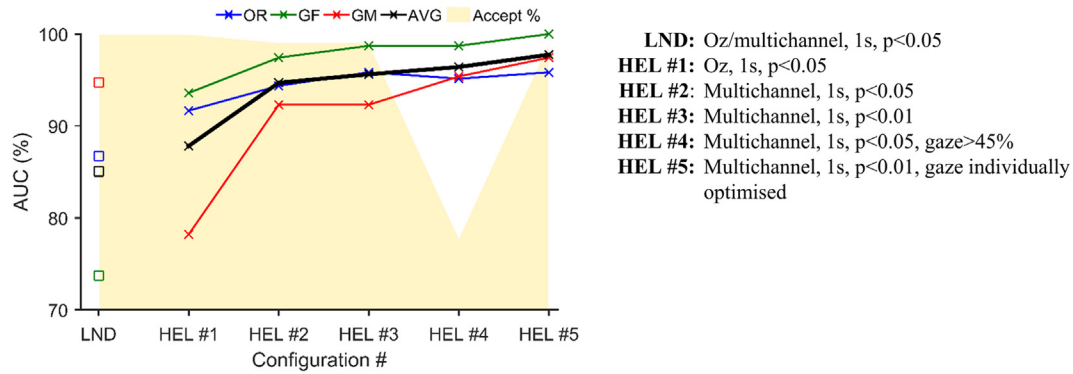
Fig. 3. Optimization of the gaze quality threshold. **A** To find the optimal threshold level for the gaze quality in the epoch selection, characteristics of the visual responses were studied as a function of variable gaze threshold ranging from 0% to 95%. Average response rates were then calculated for each step using the T^2circ method (1-s-long epochs, EEG referenced to frontal average, only posterior channels included, $p < 0.01$). Recordings during OR, GF, and GM stimuli were processed separately, and their results are depicted with blue, green, and red, respectively, with the black graph denoting the average of all stimuli. A lower limit of 60 epochs was applied for the statistical analyses. If the number of the epochs was lower than 60 (after the gaze quality thresholding), that recording was removed from the calculations. The proportion of accepted recordings for each gaze quality threshold step is represented by the brown area in the background. **B** The figure shows the incremental loss of EEG epoch quantity when the gaze quality threshold is raised. The distributions of available epochs in all the recordings are presented as boxplots depicting the medians (circles) and interquartile ranges (IQRs; boxes), with the whiskers showing the total range, neglecting the outliers (maximum length 1.5× IQR from the box edge). **C** Figures exemplify the influence of the eye tracker data on the response analysis by presenting the highest T^2circ values for each recording (N = 39) as a function of analyzed epochs quantity with and without the use of fixed 45% gaze quality threshold. OR, GF, and GM recordings are marked with blue circles, green squares, and red triangles, respectively, but the associated trend lines are fitted (in a least-squares sense) for all stimulus types pooled together. Although the slopes of the trend lines are positive in both figures, the use of the gaze threshold produces a stronger increase per epoch, indicating an improvement in the detection algorithm efficiency. A comparative analysis that includes other gaze quality thresholds is presented in Supplementary Fig. A.4. **D** The bar graph depicts the best thresholds for each stimulation session selected by maximizing the T^2circ value in the recordings (epoch number criterion of >60 applied if possible). OR, GF, and GM recordings were processed separately and are depicted with blue, green, and red bars, respectively, with the dotted lines denoting the corresponding averages. The black slashed line represents the mean of all the recordings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Our topographical analyses showed that the introduction of multichannel data was essential for the detection of GM responses that were spread over a broader postcentral area compared to the more restricted OR and GF responses. These findings are compatible with those of prior work in newborns and adults (Braddick et al., 2003, 2005; Wattam-Bell et al., 2010), which suggest at least

partly different cortical generator areas for the three responses. The OR responses are considered to arise in the V1 area and provide an early shape-selective activity (Braddick et al., 2005) to serve as the basis for object- and pattern-selectivity later in the striate and temporal lobe areas. The GF and GM stimuli are considered to reflect global processing in the ventral and dorsal visual

LND: Oz/multichannel, 1s, p<0.05
HEL #1: Oz, 1s, p<0.05
HEL #2: Multichannel, 1s, p<0.05
HEL #3: Multichannel, 1s, p<0.01
HEL #4: Multichannel, 1s, p<0.05, gaze>45%
HEL #5: Multichannel, 1s, p<0.01, gaze individually optimised

Fig. 4. Evaluation of the response detection performance. Factors affecting response detection were evaluated by comparing "area under receiver operating characteristic curve" (AUC) values calculated from the sensitivity and specificity figures from five different test configuration (HEL #1–5). The corresponding AUCs from the reference works (Braddick et al., 2005; Wattam-Bell et al., 2010; FPR information not available, but presumed to be equal to our findings) are presented leftmost (LND). Proceeding to the right, HEL #1 analysis was performed with only Oz channel (1 s epoch length and p < 0.05 T²circ significance level). Next, HEL #2 shows how the detection improves when all posterior EEG channels are included. In HEL #3, a tighter T²circ criterion of p < 0.01 is adopted, which eliminates all false-positive findings. HEL #4 adds fixation-based epoch segmentation using a fixed gaze quality threshold of 45%. Finally, in HEL #5 we used gaze thresholds that were adjusted to each recording individually (details in the Fig. 3D). The proportion of accepted recordings with each test configuration is represented by the brown area in the background. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
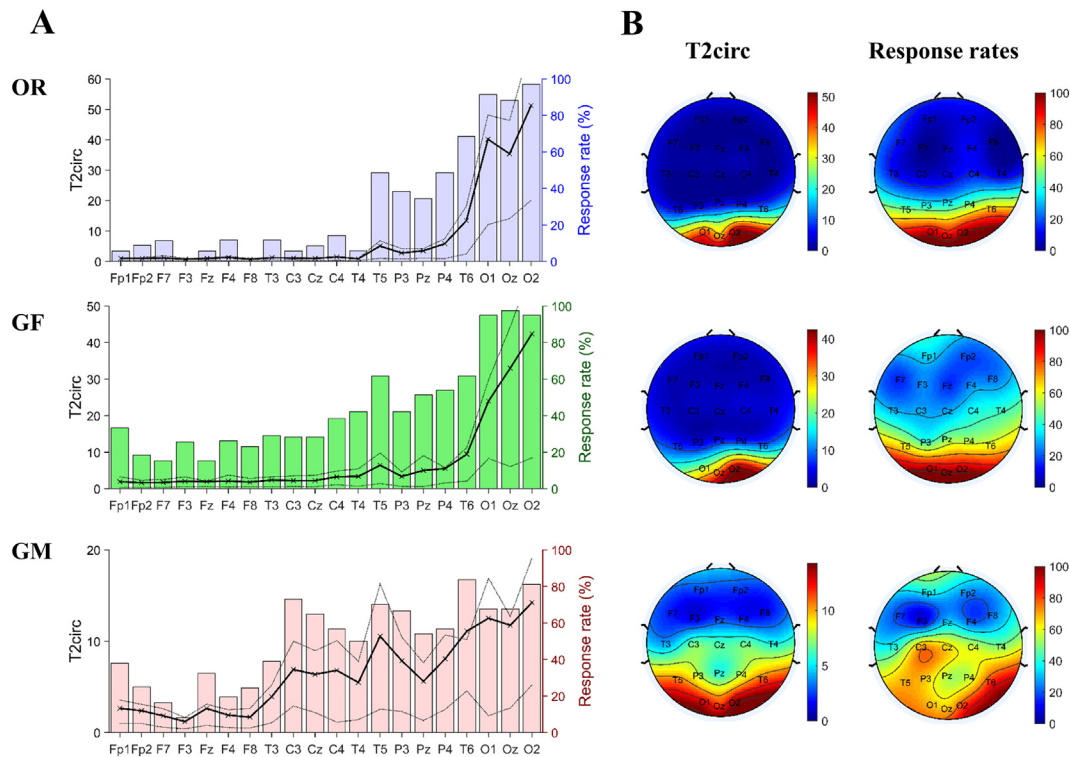
Fig. 5. Response topography. **A** Channel-specific response magnitudes (black line; left-sided y-axis) were calculated by averaging the T²circ values from the subset of the study population that yielded significant responses (p < 0.05; frontal average reference; gaze threshold individually adjusted). The inter-subject variability of the response strength at each channel is characterized by the interquartile ranges of measurements, which are depicted as dotted lines in the figures. The colored bars in the background show the average response detection rates at the specific channel (p < 0.05 without FDR correction; right-sided y-axis). **B** Group averages of channel-specific average T²circ values (left) and response rates (right) are interpolated to fit into a standard infant head model (Delorme and Makeig, 2004) to yield topographic maps of the visual responses in 2D views. Note that the colorbars of the 2D maps are scaled independently, matching the maximum T²circ value in each map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

streams (Braddick et al., 2000), respectively. More recent work comparing infant and adult responses (Wattam-Bell et al., 2010) showed that the spatial organization of response topography is still developing in infants; hence, it may not be essential to focus on detailed spatial analysis in the context of the present study. Our topographic data clearly support the presence of different spatial topographies, and this information is useful for designing optimal electrode montages. For instance, the widespread GM responses

are inadequately recorded when using a grand average reference, while it can be well presented when using reference at frontal or central areas. Conversely, a simplified recording setup using only occipital electrodes and a central or frontal reference (cf. Braddick et al., 2005) could be sufficient for the detection of occipitally dominant OR and GF responses. Indeed, the use of multiple electrodes, a natural by-product of EEG recording, is helpful for achieving better response detection rates than in earlier literature.

4.1. Clinical practicalities of the ERVS method

The stimulus presentation and the offline response analysis were implemented as semi-automated applications, which allows to perform the study with less technical expertise. In practice, our recordings were performed by our clinical neurophysiology technicians after less than an hour of training.

The cost of eye tracking devices has come down to the level of that of EEG accessories, and they are readily available in many consumer applications. This enables fast and simple integration with any external display that is capable of high-quality visual stimulations.

From the technical point of view, the most challenging part in our experimental setup was the synchronization between stimulus presentation, eye tracker, and EEG recorder. This was successfully solved by using a commercial product that captures luminance change directly from the screen; however, future implementation could build the synchronization directly into the evoked response-recording system.

4.2. Barriers to entry into clinical diagnostics

Applicability of the developed methods at the individual level for clinical evaluation would still require validation, which is conventionally performed by calculating the statistical performance indicators of sensitivity and specificity. The orthodox derivation of these figures would require "ground truths" of the studied features of interest. If this information was available, the p-value criterion could be adjusted to optimize the receiver operating characteristics of the classification, i.e., balance between true and false findings.

In this study, all measured infants were, based on their age and history, principally likely to yield responses. Thus, the observed detection rates, which were high during stimulation and low during baseline segments, hold high promise for the ERVS method's potential clinical use in differential diagnostics at an individual level. However, because of the challenges mentioned above, the validation will probably come from the test of time by providing a perceived added value to studies seeking early developmental biomarkers. Validation in this manner will be based on larger scale recruitments of both typically and atypically developing infants, which is already underway in our laboratory.

5. Conclusions

In this work, we presented an eye tracker-assisted method for the detection of visual responses to complex stimuli in the EEG of infants. Changes to the experimental setup and analysis protocol and the incorporation of information from an eye tracker resulted in a statistically significant ($p < 0.01$) response to presented visual stimuli with a negligible amount of false positives in nearly all measured infants. The output of the work is a test protocol suitable for the clinical environment where individual diagnostics are required.

Conflict of interest

None.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cnp.2017.03.002.

References

Aarnoudse-Moens, C.S., Weisglas-Kuperus, N., van Goudoever, J.B., Oosterlaan, J., 2009. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. Pediatrics 124, 717–728.

Atkinson, J., Anker, S., Rae, S., Weeks, F., Braddick, O.J., Rennie, J., 2002. Cortical visual evoked potentials in very low birthweight premature infants. Arch. Dis. Child. Fetal Neonatal Ed. 86, F28–F31.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Binkofski, F., Buxbaum, L.J., 2013. Two action systems in the human brain. Brain Lang. 127, 222–229.

Braddick, O.J., Atkinson, J., 2011. Development of human visual function. Vision. Res. 51, 1588–1609.

Braddick, O.J., Wattam-Bell, J., Atkinson, J., 1986. Orientation-specific cortical responses develop in early infancy. Nature 320, 617–619.

Braddick, O.J., O'Brien, J.M.D., Bell, W.J., Atkinson, J., Turner, R., 2000. Form and motion coherence activate independent, but not dorsal/ventral segregated, networks in the human brain. Curr. Biol. 10, 731–734.

Braddick, O.J., Atkinson, J., Wattam-Bell, J., 2003. Normal and anomalous development of visual motion processing: motion coherence and 'dorsal-stream vulnerability'. Neuropsychologia 41, 1769–1784.

Braddick, O.J., Birtles, D., Wattam-Bell, J., Atkinson, J., 2005. Motion- and orientation-specific cortical responses in infancy. Vision. Res. 45, 3169–3179.

Braddick, O.J., Atkinson, J., Newman, E., Akshoomoff, N., Kuperman, J.M., Bartsch, H., et al., 2016. Global visual motion sensitivity: associations with parietal area and children's mathematical cognition. J. Cogn. Neurosci. 28, 1897–1908.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.

Cloutman, L.L., 2013. Interaction between dorsal and ventral processing streams: where, when and how? Brain Lang. 127, 251–263.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. J. Neurosci. Methods 134, 9–21.

Freeman, W.J., Holmes, M.D., Burke, B.C., Vanhatalo, S., 2003. Spatial spectra of scalp EEG and EMG from awake humans. Clin. Neurophysiol. 114, 1053–1068.

Grinter, E.J., Maybery, M.T., Badcock, D.R., 2010. Vision in developmental disorders: is there a dorsal stream deficit? Brain Res. Bull. 82, 147–160.

Hotelling, H., 1931. The generalization of student's ratio. Ann. Math. Stat. 2, 360–378.

Johnson, S., Marlow, N., 2016. Early and long-term outcome of infants born extremely preterm. Arch. Dis. Child. http://dx.doi.org/10.1136/archdischild-2015-309581.

Johnson, S., Strauss, V., Gilmore, C., Jaekel, J., Marlow, N., Wolke, D., 2016. Learning disabilities among extremely preterm children without neurosensory impairment: comorbidity, neuropsychological profiles and scholastic outcomes. Early Hum. Dev. 103, 69–75.

Lee, J., Birtles, D., Wattam-Bell, J., Atkinson, J., Braddick, O.J., 2012. Orientation-reversal VEP: comparison of phase and peak latencies in adults and infants. Vis. Res. 63, 50–57.

Luck, S.J., 2005. An Introduction to the Event-Related Potential Technique. MIT Press, Cambridge, Mass.

Mann, H.B., Whitney, D.R., 1946. On A Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. Ann. Math. Stat. 18, 50–60.

Merchant, N., Azzopardi, D., 2015. Early predictors of outcome in infants treated with hypothermia for hypoxic-ischaemic encephalopathy. Dev. Med. Child Neurol. 57, 8–16.

Mercuri, E., Atkinson, J., Braddick, O.J., Anker, S., Nokes, L., Cowan, F., et al., 1996. Visual function and perinatal focal cerebral infarction. Arch. Dis. Child. Fetal Neonatal Ed. 75, 76–81.

Milner, A.D., 2012. Is visual processing in the dorsal stream accessible to consciousness? Proc. Biol. Sci. 279, 2289–2298.

Norcia, A.M., Appelbaum, L.G., Ales, J.M., Cottereau, B.R., Rossion, B., 2015. The steady-state visual evoked potential in vision research: a review. J. Vis. 15, 1–46.

O'Brien, J., Spencer, J., Atkinson, J., Braddick, O.J., Wattam-Bell, J., 2002. Form and motion coherence processing in dyspraxia: evidence of a global spatial processing deficit. NeuroReport 13, 1399–1402.

Odabaee, M., Freeman, W.J., Colditz, P.B., Ramon, C., Vanhatalo, S., 2013. Spatial patterning of the neonatal EEG suggests a need for a high number of electrodes. Neuroimage 68, 229–235.

Odabaee, M., Tokariev, A., Layeghy, S., Mesbah, M., Colditz, P.B., Ramon, C., et al., 2014. Neonatal EEG at scalp is focal and implies high skull conductivity in realistic neonatal head models. Neuroimage 96, 73–80.

Odom, J.V., Bach, M., Brigell, M., Holder, G.E., McCulloch, D.L., Tormene, A.P., et al., 2010. ISCEV standard for clinical visual evoked potentials (2009 update). Doc. Ophthalmol. 120, 111–119.

Picton, T.W., Bentin, S., Berg, P., Doncin, E., Hillyard, S.A., Johnson, R., et al., 2000. Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. Psychophysiology 37, 127–152.

Picton, T.W., John, M.S., Dimitrijevic, A., Purcell, D., 2003. Human auditory steady-state responses. Int. J. Audiol. 42, 177–219.

Robertson, C.E., Thomas, C., Kravitz, D.J., Wallace, G.L., Baron-Cohen, S., Martin, A., et al., 2014. Global motion perception deficits in autism are reflected as early as primary visual cortex. Brain 137, 2588–2599.

Saigal, S., Doyle, L.W., 2008. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet 371, 261–269.

Sellier, E., Surman, G., Himmelmann, K., Andersen, G., Colver, A., Krägeloh-Mann, I., et al., 2010. Trends in prevalence of cerebral palsy in children born with a birthweight of 2,500 g or over in Europe from 1980 to 1998. Eur. J. Epidemiol. 25, 635–642.

Victor, J.D., Mast, J., 1991. A new statistic for steady-state evoked potentials. Electroencephalogr. Clin. Neurophysiol. 78, 378–388.

Wattam-Bell, J., Birtles, D., Nyström, P., von Hofsten, C., Rosander, K., Anker, S., et al., 2010. Reorganization of global form and motion processing during human visual development. Curr. Biol. 20, 411–415.