Eye contact and emotional face processing in 6-month-old infants: Advanced statistical methods applied to event-related potentials

Daniel Stahl a,*, Eugenio Parise b,c, Stefanie Hoehl b,d, Tricia Striano b,e,*

a Department of Biostatistics, Institute of Psychiatry, King’s College London, De Crespigny Park, PO Box 20, London SE5 8AF, UK
b Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
c Department of Educational Sciences, Calabria University, Italy
d Department of Psychology, Heidelberg University, Heidelberg, Germany
e Department of Psychology, Hunter College, New York, NY, USA

Received 6 October 2008; received in revised form 26 March 2009; accepted 1 April 2009

Abstract

Event-related potential (ERP) studies with infants are often limited by a small number of measurements. We introduce a weighted general linear mixed model analysis with a time-varying covariate, which allows for the efficient analysis of all available event-related potential data of infants. This method allows controlling the signal to noise ratio effect on averaged ERP estimates due to small and varying numbers of trials. The method enables analyzing ERP data sets of infants, which would often not be possible otherwise. We illustrate this method by analyzing an experimental study and discuss the advantages in comparison to currently used methods as well as its potential limitations. In this study, 6-month-old infants saw a face showing a neutral or an angry expression in combination with direct or averted eye gaze. We examined how the infant brain processes facial expressions and whether the direction of eye gaze has an influence on it. We focused on the infant Negative Central ERP component (Nc). The neutral expression elicited larger amplitude and peaked earlier than the angry expression. An interaction between emotion and gaze was found for Nc latency, suggesting that emotions are processed in combination with eye gaze in infancy.

Keywords: Infants; ERP; Weighted mixed model analysis; Emotional expressions; Eye gaze; Nc component

1. Introduction

The use of event-related potentials (ERPs) can shed new light on the development of facial expression processing and its underlying neural correlates. ERPs are a useful and powerful tool in infant research, providing information about cognitive development in the absence of overt behavioural responses (for a review see [1]). However, only a relatively small number of ERP studies with (especially young) infants are published. A main reason for the small number of studies in comparison with children and adult studies is the difficulty to obtain a sufficient number of infants with analyzable measurements. In this paper we present a new statistical approach, which allows to efficiently analyzing all available ERP data of infants.

The measurement of electrical activity in the cerebral cortex recorded from electrodes placed on the scalp is a widely used method in brain research, which allows for an objective analysis of cognitive functions in infants. ERPs are usually quantified by measuring the amplitude and latency of observable peaks of the signal time-locked to stimulus or response events [2]. However, due to the physiological and physical properties of the
skull, scalp and the brain the signal to noise ratio of ERP recordings is often low and waveforms differ between trials of the same stimulus. Measurements of amplitude and latency of observed peaks differ between trials and the variance will be larger the smaller the signal to noise ratio is. The most common method to reduce this resolution problem is to average a large number of trials, assuming that the ERP response is consistent to repeated stimuli of the same kind. Given enough trials, the averaging process will filter out noise that is not related to the stimulus and results in an ERP waveform related to the stimulus [3]. The number of trials needed for a precise and reliable average depends on the signal to noise ratio. If the signal of the stimulus does not change over time the signal to noise ratio increases as a function of the square root of the number of trials. In adult studies often hundreds of trials are averaged. However, the assumption of consistency of response to the stimuli may not be valid in a large number of trials due e.g. to habituation processes [4].

In experimental studies with infants similarly large numbers of trials as in adult studies are not feasible and an aggregated measurement is typically used if at least 10 artifact free trials per experimental condition are obtained [2]. However, even this rule of thumb is often problematic, since the number of valid trials per experiment varies between and within infants. In many cases infants contribute only a few trials per experimental condition. In this study we tested 53 infants of which 46 provided ERP data. Only 14 out of 46 infants or 30% of infants tested passed the criterion of at least 10 trials in at least one of four experimental conditions. In addition, only 3 out of these 14 infants passed the criterion in all four experimental conditions. Analyzing the data of 14 children would reduce the statistical power of our analysis. It could also result in seriously biased parameter estimates if data are not missing completely at random (MCAR, see [5] for a detailed discussion about missing values); an assumption which is usually not met [6]. It is conceivable that infants who do not provide a sufficient number of trials for the data analysis are on a different developmental level than infants who constitute the data set of the analysis. In this case the results would be biased (some members of the population are more likely to be included than others) and not representative for the population under investigation.

Furthermore, the standard statistical analysis of within group designs is still repeated measure ANOVA (or repeated measure MANOVA), which performs a complete cases (i.e. all infants) analysis. Analyzing our data set with standard methodology would reduce sample size further to $N = 3$, which would not allow any inference.

In this paper we suggest a new method to overcome these highlighted problems. The precision of an estimate of the amplitude or latency of the peaks of a brain signal to a stimulus depends on the signal to noise ratio. Assuming that the signal is unaffected by the repeated trials, estimates based on a smaller number of trials should be less precise, that is the variance should increase with decreasing number of trials. However, one of the assumptions underlying linear modeling techniques, such as ANOVA and regression, is that each observation provides equally precise information, i.e. the variance of the error term is constant over all values of the predictor or independent variables. A violation of the homogeneity of variance assumption will result in unbiased but inefficient parameter estimation: The standard error will be either too small (variance increases with increasing value of independent variable) or too large (variance decreases with increasing value of independent variable), which may lead to a Type I or Type 2 error. A common approach to overcome non-constant variance in the response variable is to use an inverse variance weighted regression approach [7,6]. A weighted regression analysis allows adjusting for different precision of aggregated measurements. Cases are weighted by the reciprocal of their estimated variance, thereby attributing more weight on cases with smaller estimated variances (i.e. cases with a larger number of artifact free trials).

A further complication arises if the signal to noise ratio is large because with larger number of trials more noise is averaged out. In this case, the peak amplitude will be less affected by noise and therefore decrease [3]. If the number of trials affects the observed averaged response signal, the number of trials needs to be included as a covariate to adjust for the influence of the number of trials on the averaged response. Because the signal to noise ratio decreases with the square root of the number of trials, a linear and a quadratic term of the number of trials need to be included in the model to model the non-linear relationship. Using this method, we assume that the number of trials is not determined by the experimental condition. Since we assume that different experimental conditions will elicit different ERP responses, including the number of trials in the model would bias our result if the number of trials correlated with the Nc response. This assumption needs to be evaluated by comparing the number of trials within an infant. It should be noted that the same assumption has to be made for the standard analysis of infant ERP data sets. To assess the assumption that the experimental conditions do not influence the number of trials, we performed a variance component analysis to assess the absolute consistency (or agreement) of the number of trials within infants among experimental conditions [8].

We used a weighted general linear mixed model (GLMM) analysis (also known as multilevel or hierarchical model analysis) to analyze the data set. GLMM
using maximum likelihood estimation methods allows a full case analysis and is furthermore less restricted on the type of missingness than standard repeated measurement ANOVA analysis [9]. Furthermore, GLMM permits to include time-varying covariates, such as the number of trials in each experimental condition. To reduce the problem of differences in the precision of our data we use weights that are inversely proportional to the variance at each level of the explanatory variables “number of trials”, which yields efficient parameter estimates. To assess sensitivity, we compared the results of the weighted analysis with an unweighted analysis to assess the influence of infants with a larger number of trials (i.e. larger weights) on the estimated differences between the experimental conditions.

We illustrate the use of an invariance-weighted GLMM with the number of trials as a potential covariate by analyzing an experimental study on Emotion and Gaze perception in infants. Our approach allows to analyze ERP studies with infants in a more efficient and precise way. We will describe the method in details and discuss the advantages in comparison with standard analysis as well as its limitations.

We introduce our approach based on data from a study on emotional expression and gaze processing in 6-month-old infants. Nelson and de Haan [10] found that 7-month-old infants differentiate happy from fearful facial expressions but not fearful from angry. Specifically, on fronto-central and parietal electrodes the Negative Central component (Nc) was larger for a fearful face compared to a happy one; whereas the Positive Slow Wave (PSW, related to memory update) showed the opposite pattern and was larger for the happy expression (see [11], for details about these infantile ERP components and source localization of the Nc). The authors interpreted these results as a function of familiarity with different facial expressions.

Some studies also investigated the processing of eye gaze in infants using ERP measures. For instance, Farroni and colleagues [12] found that 4-month-olds showed an enhanced N290 ERP component when looking at faces with direct eye contact compared to faces with averted gaze. In a series of infant ERP studies, direct eye contact showed to be an important social cue enhancing object processing in triadic person–object–person interactions [13,14].

In related research, 4-month-old infants showed an enhanced PSW in response to angry faces with direct gaze compared to averted gaze. No effects of gaze direction were found for happy or neutral faces [15]. In this study, happy, neutral and angry facial expressions were presented to separate groups of infants. The authors did not find significant differences between conditions at earlier latencies. By comparing angry and fearful facial expressions, Hoehl and Striano [16] found that at 7 months of age, angry faces with direct eye gaze elicit an increased Nc amplitude relative to angry faces with averted gaze. Webb et al. [17] showed that the Nc amplitude increases over time between 4 and 12 months (i.e. becomes more negative). At the same time its latency decreases. Interestingly, these changes of the Nc across age interacted with different response for the stimuli. At 4 and 6 months the Nc amplitude is more pronounced for faces compared to objects, whereas from 8 months this pattern is reverted. This pattern of results suggests greater neural activation and faster information processing occurring in the infant brain especially between 4 and 8 month of age, accompanied by changes in processing of visual stimuli, including faces.

The present study was designed to better understand emotional facial expression processing in 6-month-olds. Six-month-old infants were shown a face with a neutral or an angry expression in combination with direct or averted eye gaze. We examined how the infant brain processes facial expressions and whether the direction of eye gaze has an influence on it. Based on the results of Striano et al. [15] we expect that emotional expressions are not processed in isolation from the eye gaze. We focused our investigation on the well-characterized Nc component. This typical infant component has different characteristics as a function of the employed experimental paradigm, and its functional meaning can be related to both recognition memory and activation of attentional resources (see [11]). When stimuli were not preceded by a familiarization period and were novel to the infant, the Nc was more negative to infrequent stimuli [18]. When highly familiar stimuli were presented, the Nc was more negative for familiar stimuli compared to new ones [19,20]. Finally, when infants were previously familiarized with the stimuli before the experimental session, the Nc provided more variable results. Sometimes it was larger for novel stimuli compared to familiar ones [11], other times this difference was not found [10,23,22]. However, Richards [22] established a clear relation between Nc component and attention by relating its amplitude to heart rate deceleration measurement.

2. Experimental procedure

2.1. Participants

Forty-six infants were included in the final analysis (26 females). Average age was 5 months and 28 days (SD = 7.85 days; range from 5 months and 15 days to 6 months and 13 days). All infants were born full term (37–41 weeks) and were in the normal range for birth weight. Another seven infants were tested but excluded because they were too fussy or crying (n = 2), because of technical problems (n = 2), or because they were not at all looking at the stimuli (n = 3).
2.2. Stimuli

Stimuli used in this experiment were the same as the ones used by Striano et al. [15]. A female face showing a neutral or angry facial expression was presented with direct or left/right averted gaze (see Fig. 1).

2.3. Procedure

Infants sat on their mother’s lap in a dimly lit sound-attenuated and electrically shielded cabin. Viewing distance was 70 cm away from a 70 Hz 17-in. stimulus monitor. The experiment consisted of one blocks of 192 trials. Each trial contained a face with a neutral or angry emotional expression and direct or averted eye gaze, resulting in a $2 \times 2$ experimental design. Faces were presented on a black background and in pseudo-random order with the constraint that the same condition was not presented three times consecutively. Each trial was preceded by a small cross (as a central fixation attractor), presented at the center of the screen for 500 ms, followed by each stimulus presented on the screen for 1 s. Between the presentations of the trials, the screen was blank for a random period of between 1000 ms and 1200 ms. Whenever an infant became fussy or stopped looking at the stimuli, the experimenter gave the infant a short break. Electroencephalogram (EEG) was recorded continuously and infant behaviour was video-recorded throughout the session for offline trial-by-trial editing of the EEG in order to ensure that the infant was looking at the screen for all trials included in later analysis.

2.4. EEG recording and analysis

EEG was recorded with Ag–AgCl electrodes from 23 scalp locations of the 10–20 system, referenced to the vertex (Cz). Data were amplified via a Twente Medical Systems 32-channel REFA amplifier. Horizontal and vertical electrooculogram were recorded bipolarly. Sampling rate was set at 250 Hz. EEG data were re-referenced offline to the linked mastoids. Offline bandpass filter was set from 0.3 to 20 Hz.

The EEG recordings were segmented into epochs of waveform that comprised a 200 ms baseline and 1000 ms of the image displaying the object. For the elimination of electrical artifacts caused by eye and body movements, the EEG data were rejected off-line.
whenever the standard deviation within a 200 ms gliding window exceeded 50 \mu V at any electrode and 80 \mu V for EOG ([15,16]). Data were also visually edited offline for artifacts and to ensure that included trials were those where the infant was attending to the screen.

To assess differences in the Nc, a time window from 220 to 700 ms was chosen on frontal and central channels. For each included subject the minimum amplitude within this window was selected in frontal-central left (F3, C3), frontal-central (Fz, Cz) and frontal-central right (F4, C4) locations. We assessed differences both for the latency and amplitude of the Nc peak within the mentioned time window.

2.5. Statistical analysis

2.5.1. Intraclass correlation

We used a variance component analysis (two-way ANOVA) to estimate the intraclass correlation coefficient for absolute consistency or agreement (ICC2(A,1), Fleiss and Shrout [8]) between the four experimental conditions:

\[
\text{ICC for absolute agreement : } \frac{\sigma^2_{\text{infants}}}{\sigma^2_{\text{infants}} + \sigma^2_{\text{condition}} + \sigma^2_{\text{error}}}
\]

The ICC measures if experimental conditions assign the same absolute number of trials within an infant. An ICC of 1 would mean that the variance is fully explained by differences between infants and not by different experimental conditions.

2.5.2. Weighted analysis

The aggregated measures are based on a mean number of 6.5 trials per condition (SD = 4.3, range 1–23). Only 13% of the observations are based on 10 or more trials per condition. In a first step we plotted the infants’ averaged responses of Nc latency and amplitude against the number of trials for each experimental condition. If the number of trials influences the precision of the mean response rate, we would expect an increase in variance with decreasing number of trials. Measurements calculated from a smaller number of trials are less precise than measurements based on larger number of trials. A common approach to overcome non-constant variance in the response variable is to use an inverse variance weighted regression approach by weighting the contribution of a measurement with the inverse of the estimated variance for the number of trials of this value [6,7]. This method ensures that cases which are based on a smaller number of trials and thus higher variance (i.e. those with less precision) will count less and cases based on a larger number of trials (i.e. those cases with more precision) count more in the estimation of the parameters in a model. The weights were estimated using the power of the independent variable “number of trials”, which maximizes the likelihood of the dependent variable “latency” or “amplitude”. The inverse of the power is then used to weight the cases in the final analysis. This method assumes that a decrease in variance is characterized by the variance of the dependent variable increasing exponentially at some power function of the “number of trials”. The estimated weights are inversely related to the variance at each level of the explanatory variable. Using these weights in a GLMM analysis will result in more efficient and more precise parameter estimates. We estimated the weights using “Weight Estimation” in SPSS’s regression module. To ensure that the weights improved the fit of the model we assessed plots of weighted predicted values versus weighted predicted residuals. For both weighted analyses, variance was constant over the number of trials, while a decrease of variance over the number of trials was noticeable with the respective unweighted analyses.

2.5.3. Adjusting for the number of trials

If the number of trials influences the parameter estimate for the response we would observe a consistent change over the number of trials used for averaging. In this case the number of trials needs to be included as a covariate in the model to adjust for different number of trials especially between conditions.

We included the number of trials as a condition-varying covariate (the number of trials for an infant differed between the experimental conditions) in the analysis of the Nc amplitude because we observed a decrease in the averaged amplitude with increasing number of trials used for the averaging (see Section 3). Because the observed relationship was non-linear, we included number of trials as a quadratic model. To ensure that the quadratic model adequately describes the relationship between the number of trials and responses, we plotted the estimated curve (based on the parameter estimates of the GLMM analysis) against the observed averaged responses.

2.5.4. General linear mixed model analyses

General linear mixed models (GLMMs) are an extension of repeated measurement analyses of variances [8]. Unlike repeated measurement ANOVAs, they allow to analyze complete data sets in case of missing values and enable to weight each observation individually. A GLMM analysis uses maximum likelihood estimation methods, which are less restrictive in the missingness assumptions than least squares estimation methods used in standard repeated measurement ANOVAs. Repeated measurement ANOVA gives only unbiased estimates if data are missing completely at random, while GLMM allow missingness as long as the model is correctly specified and the reasons for the missingness of the value does not depend on the value being missing (missing at random (MAR), see [5]). Furthermore, mixed models allow including time- or condition-varying covariates:
Unlike in repeated measure ANOVA, the value of a covariate (number of trials) is allowed to vary between conditions within an infant.

2.5.5. Experimental design

Each infant was observed under two experimental conditions, Emotion and Gaze, in a crossed design. Under each possible combination measurements were taken at three different locations (left, central and right) and within each location two measurements, electrodes, were taken (respectively: F3, C3; FZ, CZ; F4, and C4). Therefore for each infant we obtained $2 \times 2 \times 3 \times 2 = 24$ measurements of the Nc component both for latency and amplitude.

To analyze the responses we used general linear mixed models with the two experimental conditions “Emotion” and “Gaze” and the interaction between Emotion and Gaze as fixed factor. Location (left, central, and right) and Electrodes (frontal or central electrodes) were entered as two additional fixed factors. To account for the repeated observation of the same individual, “infant id” was included as a random factor in the model. Additionally, we tested all possible two-way interactions among fixed factors as well as interactions between the random factor “infant id” and the fixed factors “Gaze” and “Emotion”. Based on Akaike’s information criteria (AIC) and model diagnostic fits [8], an additional interaction between Emotion and Location was selected in the final model of the latency analysis. All other interactions did not improve the explanatory power of the models and were thus excluded.

To assess how robust the results are to assumptions about the data and the method that was used we repeated the analyses by rerunning the final model without weights and number of trials as a covariate as well as with observations which are based on at least 10 trials.

Parameter estimates based on models which included the number of trials as a quadric trend were evaluated at seven trials (rounded mean number of trials).

The assumptions of the general linear mixed model analysis were assessed by standard visual inspections of the residuals and random effect estimates (see [6,8] for details). All analyses were done using SPSS 15.0 for Windows [21].

3. Results

3.1. Analysis of “number of trials” across groups: intraclass correlation

The mean number of trials did not differ between the four experimental conditions $F(3, 135) = 1.69, p = 0.17$.

The ICC for absolute agreement for the number of trials among the four experimental conditions was 0.89. That is, 89% of the variance of the number of trials is explained by between-infant differences and only little variation (11%) was observed within an infant. Of this variation within infants less than 1% of the variance could be attributed to between experimental condition differences (Variance components: infant: 17.52, experimental condition: 0.03, error: 2.15). The number of trials is mainly an infant characteristic.

3.2. Analysis of Nc component: latency

3.2.1. Results for full data set – weighted GLMM: latency

Fig. 2a shows the scatter plot of the number of trials against the averaged responses for each infant under each experimental condition. For simplicity the responses of the different electrodes of each infant are shown within each experimental condition. Plotting the response for each electrode type separately resulted in similar relationships. The plots show only a minor decrease in variance with increasing number of trials used for the average. Consequently a fairly small weight was estimated for the weighted GLMM analysis. There were no changes in mean response in dependency of the number of trials between 1 and 15 trials. The decrease in the lowess curve afterwards is caused by the response of one infant and can be explained by chance variation. The overall pattern of the trajectories is similar for each experimental condition.

The GLMM analysis was weighted by the inverse of the (number of trials)$^{-0.12}$. Because the number of trials seems not to influence the mean response, we did not include the number of trials in our main model. There was a significant main effect for Emotion (Table 1a: model 3). Overall, infants showed a shorter latency in the “neutral” condition compared to the “angry” condition (Marginal means: 0.461 (SE = 0.01) versus 0.477 (SE = 0.01)). However, there was a significant interaction between Emotion and Gaze (Fig. 3). The difference in latencies between the “toward” and “away” condition was larger in the “angry” compared to the “neutral” condition (mean difference of change score: 0.041 (95% C.I.: 0.015–0.067)). Pairwise comparisons revealed that infants showed in the “angry” condition a significant longer latency in the “away” condition compared to the “toward” condition ($p = 0.022$). The opposite was observed in the “neutral” condition: infants showed shorter latencies in the “away” condition compared to the “toward” condition ($p = 0.036$). Furthermore, in the “away” condition, infants showed a significant longer latency in the “angry” condition compared to the “neutral” condition ($p < 0.0001$), while there was no significant difference in the “toward” condition ($p = 0.479$).

Additionally, the GLMM analysis revealed a significant main effect for Electrode: infants showed a shorter latency in the central area compared to the...
There was a significant interaction between Emotion and Location. Overall, infants showed a longer latency in the "angry" condition compared to the "neutral" condition in the right location (marginal mean difference 0.035 (SE = 0.011), \(p = 0.002\)), while no significant differences were observed in central (\(p = 0.353\)) and left locations (\(p = 0.720\)). There were no significant main effects for Gaze and Location.

3.2.2. Control analysis I – weighted GLMM with number of trials: latency

The results remained virtually unchanged if the number of trials was included as a quadratic model (Table 1a: model 1).
Table 1

The results of the general linear mixed model analysis for (a) Nc latency and (b) Nc amplitude measurements. For each measure the results of the linear model analysis (type III F-Tests using Satterwaites approximations) are presented (degrees of freedom (df), F-value and p-value). The marginal means and 95% confidence intervals for each level of the main effects, Emotion and Gaze, are presented together with the estimated mean difference between the two levels within a factor. Furthermore, marginal means and 95% confidence intervals (95% C.I.) relevant to the interaction between Emotion and Gaze are presented together with interaction contrast (mean difference of change score: i.e. the difference in the change score between “Angry” and “Neutral” condition between “Toward” and “Away” condition). Parameter estimates based on models, which included the number of trials as a quadric trend (models 1 and 2), were evaluated at number of trials = 7.

(a) Latency: Model 1: Weighted + Trials Model 2: Trials only Model 3: Weighted Model 4: without weights and trials Model 5: 10+ trials

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1: Weighted + Trials</th>
<th>Model 2: Trials only</th>
<th>Model 3: Weighted</th>
<th>Model 4: without weights and trials</th>
<th>Model 5: 10+ trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df’s</td>
<td>F-Value</td>
<td>p-Value</td>
<td>df’s</td>
<td>F-Value</td>
</tr>
<tr>
<td><strong>Emotion</strong></td>
<td>1,1023.7</td>
<td>4.85</td>
<td>0.028</td>
<td>1,1023.2</td>
<td>3.94</td>
</tr>
<tr>
<td><strong>Gaze</strong></td>
<td>1,1029.2</td>
<td>0.97</td>
<td>0.37</td>
<td>1,1029.1</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Emotion × Gaze</strong></td>
<td>1,1024.9</td>
<td>9.29</td>
<td>0.002</td>
<td>1,1025.1</td>
<td>10.46</td>
</tr>
<tr>
<td><strong>Electrodes</strong></td>
<td>2,1011.8</td>
<td>0.04</td>
<td>0.94</td>
<td>2,1012.3</td>
<td>7.8</td>
</tr>
<tr>
<td><strong>Emotion × Location</strong></td>
<td>2,1011.8</td>
<td>3.1</td>
<td>0.046</td>
<td>2,1012.3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Number of trials</strong></td>
<td>1,123</td>
<td>3.1</td>
<td>0.08</td>
<td>1,126.3</td>
<td>2.94</td>
</tr>
<tr>
<td><strong>Number of trials 2</strong></td>
<td>1,95.1</td>
<td>4.97</td>
<td>0.028</td>
<td>1,95.1</td>
<td>4.97</td>
</tr>
</tbody>
</table>

(b) Amplitude: Model 1: Weighted + Trials Model 2: Trials only Model 3: Weighted Model 4: without weights and trials Model 5: 10+ trials

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1: Weighted + Trials</th>
<th>Model 2: Trials only</th>
<th>Model 3: Weighted</th>
<th>Model 4: without weights and trials</th>
<th>Model 5: 10+ trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df’s</td>
<td>F-Value</td>
<td>p-Value</td>
<td>df’s</td>
<td>F-Value</td>
</tr>
<tr>
<td><strong>Emotion</strong></td>
<td>1,1037.9</td>
<td>14.97</td>
<td>0.0001</td>
<td>1,1022.5</td>
<td>20.91</td>
</tr>
<tr>
<td><strong>Gaze</strong></td>
<td>1,1023.4</td>
<td>3.41</td>
<td>0.065</td>
<td>1,1027.6</td>
<td>1.59</td>
</tr>
<tr>
<td><strong>Emotion × Gaze</strong></td>
<td>1,1021.4</td>
<td>3.57</td>
<td>0.05</td>
<td>1,1024.7</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Electrodes</strong></td>
<td>1,1006.8</td>
<td>3.88</td>
<td>0.049</td>
<td>1,1021.1</td>
<td>1.18</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>2,1006.8</td>
<td>3.09</td>
<td>&lt;0.0001</td>
<td>2,171.6</td>
<td>5.35</td>
</tr>
<tr>
<td><strong>Emotion × Location</strong></td>
<td>2,1006.8</td>
<td>0.4</td>
<td>0.67</td>
<td>2,1021.2</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>Number of trials</strong></td>
<td>1,126.1</td>
<td>32.5</td>
<td>&lt;0.0001</td>
<td>1,196.2</td>
<td>26.72</td>
</tr>
<tr>
<td><strong>Number of trials 2</strong></td>
<td>1,128.2</td>
<td>11.41</td>
<td>0.0009</td>
<td>1,142.6</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Factor Level Estimate (95% C.I.) Estimate (95% C.I.) Estimate (95% C.I.) Estimate (95% C.I.) Estimate (95% C.I.)

**Emotion**

**Angry**

Gaze: 
- **Angry-Away**: 24.27 (23.05, 25.49) 27.09 (25.81, 28.36) 1.24 (0.82, 1.66) 0.61 (0.29, 0.94) 0.99 (0.62, 1.36)

Mean difference (A-N): 2.74 (2.65, 2.83) 25.21 (24.37, 26.05) 27.31 (26.92, 27.70) 23.8 (23.4, 24.18) 31.49 (31.17, 31.81)

Neutral-Away: 26.72 (26.06, 27.38) 27.33 (26.97, 27.69) 23.09 (22.73, 23.46) 22.93 (22.55, 23.31) 27.51 (27.13, 27.89)

Neutral-Toward: 28.49 (27.62, 29.36) 29.26 (28.91, 30.61) 29.07 (28.62, 30.52) 23.09 (22.75, 23.43) 30.57 (30.19, 31.05)

**Neutral**

Gaze: 
- **Neutral-Away**: 26.72 (26.06, 27.38) 27.33 (26.97, 27.69) 23.09 (22.73, 23.46) 22.93 (22.55, 23.31) 27.51 (27.13, 27.89)

Mean difference (A-N): 2.74 (2.65, 2.83) 25.21 (24.37, 26.05) 27.31 (26.92, 27.70) 23.8 (23.4, 24.18) 31.49 (31.17, 31.81)

Neutral-Away: 26.72 (26.06, 27.38) 27.33 (26.97, 27.69) 23.09 (22.73, 23.46) 22.93 (22.55, 23.31) 27.51 (27.13, 27.89)

Neutral-Toward: 28.49 (27.62, 29.36) 29.26 (28.91, 30.61) 29.07 (28.62, 30.52) 23.09 (22.75, 23.43) 30.57 (30.19, 31.05)

**Neutral-Toward**

Gaze: 
- **Neutral-Away**: 26.72 (26.06, 27.38) 27.33 (26.97, 27.69) 23.09 (22.73, 23.46) 22.93 (22.55, 23.31) 27.51 (27.13, 27.89)

Mean difference (A-N): 2.74 (2.65, 2.83) 25.21 (24.37, 26.05) 27.31 (26.92, 27.70) 23.8 (23.4, 24.18) 31.49 (31.17, 31.81)

Neutral-Away: 26.72 (26.06, 27.38) 27.33 (26.97, 27.69) 23.09 (22.73, 23.46) 22.93 (22.55, 23.31) 27.51 (27.13, 27.89)

Neutral-Toward: 28.49 (27.62, 29.36) 29.26 (28.91, 30.61) 29.07 (28.62, 30.52) 23.09 (22.75, 23.43) 30.57 (30.19, 31.05)

**Interaction contrast**

0.83 (-0.83, 2.49) 1.43 (0.25, 1.81) 2.03 (1.24, 2.82) 0.83 (-0.83, 2.49) 2.03 (1.24, 2.82)

**Model 1**: weighted analysis of all trials with number of trials as a covariate included, model 2: unweighted analysis of all trials with number of trials as a covariate included, model 3: weighted analysis of all trials with number of trials as a covariate included, model 4: unweighted analysis of all trials without trials as a covariate included, and model 5: conventional analysis using observation based on 10 or more trials only (N = 19).

3.3. Analysis of the Nc: amplitude

3.3.1. Results for full data set – weighted GLMM: amplitude

Fig. 2b shows the scatter plot of the number of trials against the averaged response for each infant under each emotion condition.

3.2.3. Control analysis II – unweighted GLMM: latency

The GLMM analysis was weighted by the inverse of the (number of trials)^{-0.12}, a fairly small weight. Consequently, rerunning the GLMM analysis without the weighting changed the results only marginally (Table 1a: with trials as covariate included: model 2, without trials as covariate included: model 4).

3.2.4. Control analysis III – 10+ trials GLMM: latency

In a further control analysis we ran an unweighted GLMM using observations based on 10 or more trials only (Table 1a: model 5). The analysis included 14 infants with altogether 186 observations and resulted in a similar significant interaction between Gaze and Emotion. Again, pairwise comparisons revealed that in the “angry” condition infants showed a significant longer latency in the “away” condition compared to the “toward” condition (p < 0.007), while in the “neutral” condition infants showed a shorter latency in the “away” condition compared to the “toward” condition (p < 0.002).

In the “away” condition, infants showed again a significant longer latency in the “angry” condition compared to the “neutral” condition (p < 0.0001). Unlike in the analysis based on all observations there was a significant difference between the “angry” and “neutral” condition within the “toward” condition. Infants showed a significant longer latency in the “neutral” condition compared to the “angry” condition (p = 0.029). However, a similar pattern was observed in the analysis of the full data set (see Fig. 3 and Table 1). All other main effects and interactions were not significant (all p-values > 0.163).
experimental condition. For simplicity the responses of the different electrodes of each infant are shown within each experimental condition. Plotting the response for each electrode type separately resulted in similar relationships. The plots show a decrease of variance with increasing number of trials used for the average. Consequently, a fairly strong weight was estimated for the weighted GLMM analysis. Furthermore, a consistent non-linear decrease of the mean response can be observed. The decrease continues beyond 10 trials and the pattern is similar between each experimental condition. The number of trials used for averaging affects the mean observed amplitude and the number of trials was included as a quadratic model in the analysis.

The GLMM analysis was weighted by the inverse of the (number of trials)^{-1.1}. There was a significant main effect of Emotion (Table 1b: model 1). Independent of Gaze condition, infants showed a larger (i.e. more negative) amplitude in the “neutral” condition compared to the “angry” condition (mean difference in amplitude: 2.86, 95% C.I.: 1.41–4.31), Fig. 4). There were no significant main effects for Gaze or an interaction between Emotion and Gaze.

Amplitude was larger in frontal compared to central electrodes and there was a significant main effect of Location. Infants showed a larger amplitude in the central electrodes compared to left (p < 0.0001) and right electrode (p = 0.001). The amplitude in left and right electrodes did not differ significantly (p = 0.126) (marginal means: left: −24.25 (SE = 1.65), center: −28.66 (SE = 1.65) and right electrodes: −25.60 (SE = 1.65)). The parameters estimates for the quadratic model of the number of trials were b1 = 3.51 (SE = 0.615, p < 0.0001) for the linear term and b2 = −0.10 (SE = 0.031, p = 0.001) for the quadratic term. The point at which the quadratic equation levels off is at 17.5 trials. A plot of the estimated curve against the observed averaged amplitude suggests that the quadratic model describes the relationship between the number of trials and mean amplitude well.

3.3.2. Control analysis I – unweighted GLMM with number of trials as covariate: amplitude

The GLMM analysis was weighted by the inverse of the (number of trials)^{-1.1}, a fairly strong weight. However, rerunning the GLMM analysis without the weighting did not change the main results (Table 1b: model 2). There was a significant main effect for Emotion, and the amplitude was larger in “neutral” compared to “angry” condition (mean difference: 4.77 (95% C.I.: 2.72–6.81)). Again there was a significant main effect for Location showing the same pattern as in the main analysis. All other main effects and interactions were not significant (all p’s > 0.24).

3.3.3. Control analysis II – unweighted GLMM without number of trials as covariate: amplitude

Rerunning the GLMM analysis without the weighting and the number of trials did not change the main results (Table 1b: model 4). There was a significant main effect for Emotion, and the amplitude was larger in “neutral” compared to “angry” condition (mean difference: 4.00 (95% C.I.: 1.93–6.07)). Again there was a significant main effect for Location showing the same pattern as in the main analysis. All other main effects and interactions were not significant (all p’s > 0.29).

3.3.4. Control analysis III – 10+ trials GLMM: amplitude

In a further control analysis we ran an unweighted GLMM using observations based on 10 or more trials only. The analysis included 14 infants with 186 observations and resulted in a similar significant main effect of Emotion while there was no significant main effect for Gaze or an interaction between Gaze and Emotion (Table 1b: model 5). Again, there was a significant main effect for Location and a significant effect of Electrode. The general pattern of the amplitudes among the different locations did not change.

4. Discussion

We introduced a new statistical method to analyze a within-subjects experimental design with four different conditions without missing any data and to elaborate ERP data even with a small number of artifact free trials. Our analysis of the Nc component showed that 6-month-olds differentiated between neutral and angry facial expressions and that these emotional expressions interacted with eye gaze perception, which suggests that emotional expressions are not processed in isolation from the eye gaze. Specifically, the Nc showed larger amplitude for the neutral expression and also peaked earlier for neutral than for angry faces. Furthermore,
in the gaze away condition the neutral face peaked earlier than both neutral in the direct gaze and angry in the averted eye gaze condition. Finally, the angry face peaked earlier in the direct gaze than in the averted eye gaze condition. Only in the analysis using the observations based on 10 or more trials the angry face peaked earlier than the neutral face within the direct eye gaze condition.

In other studies by de Haan and Nelson [19,20], 6-month-old infants showed a larger Nc component for highly familiar stimuli compared to the novel ones both when human faces and objects were used as stimuli. Infants of social middleclass families as the ones recruited for this experiment are unlikely exposed frequently to negative facial expressions in the first half of postnatal life [24]. Thus, the more negative amplitude of the Nc for the neutral face can be interpreted as a recognition memory effect, because the neutral facial expression was more familiar than the angry one. Interestingly, our results also showed a shorter latency of the Nc component for the neutral face compared to the angry face, which was not found in the studies by de Haan and Nelson [19,20]. This finding may indicate that the neutral expression was easier to process relative to the more unfamiliar angry face.

Our results add new insight to the growing body of literature addressing eye gaze perception and emotional facial expression processing. In fact they seem to contradict previous findings suggesting that amplitude of the Nc is related to the threat-related signal value of an emotional facial expression. For instance, Nelson and de Haan [25] found that 7-month-olds display a more prominent Nc for a fearful face compared to a happy one, whereas no difference was found between fearful and angry. This indicates that the more threatening or negative expression elicited increased attention. Hoehl and Striano [16] found that 7-month-olds show a more negative Nc to angry faces with direct gaze compared to angry faces with averted gaze. However, Nelson and de Haan [25] and Hoehl and Striano [16] used different emotion pairs compared to us. Most importantly, the Nc component changes across age, decreasing in amplitude and in latency during the first year of life [16]. Infants in the discussed studies were older than in our study, which points to an important developmental trajectory in face processing between 6 and 7 months of age.

There was also an interaction between emotional expression and eye gaze for latency of the Nc. The Nc peaked earlier for the neutral face in the averted gaze condition than for neutral faces with direct gaze while the opposite pattern was found for angry faces. This effect is interesting when considering studies demonstrating the role of eye gaze perception in infants as a signal for external objects or events. Hoehl et al. [26] reported that a face with a neutral expression and the eyes directed toward an object or averted from it resulted in differences in the Nc component of 4-month-olds. Specifically, when the eyes were directed to the object the Nc peaked earlier. Even though no object was displayed beside the face in our stimuli, we can hypothesize that the neutral facial expression with averted gaze suggested to the infant a potential focus of interest in the surrounding environment, “preparing” the infant brain to pay attention to something beside the face stimulus. This was not found when the facial expression depicted an angry expression. In this case the condition of direct eye contact with the infant elicited a faster Nc when compared to angry faces with averted gaze. This finding may be interpreted as a rapid orientation mechanism toward a potentially threatening stimulus: an angry face with direct eye gaze.

When comparing our results to those obtained by Striano et al. [15] with the same stimuli, we conclude that 6-month-olds in our study showed an earlier differentiation in the brain processing of an angry face with direct gaze from one with the same expression and averted gaze compared to 4-month-olds in the previous study. These data show that the same facial expression is processed faster depending on eye gaze direction. However, this latter result was partially obtained in a control analysis, which may be biased towards cognitively higher developed infant and further investigations are required.

Although our analysis is based on data which are aggregated from a smaller number of trials as commonly used, our results were stable across the different analysis methods and are consistent with some of the existing literature.

First, the number of trials across experimental conditions was similar and the variance component analysis revealed that the number of trials was mainly determined by the infant identity and not by experimental condition. The observed number of trials in our study varied little within an infant but considerably between infants. Therefore, we can assume that experimental condition did not influence the number of trials and did not bias the results of the weighted analysis with number of traits as a covariate. Second, rerunning the analysis without weights did not overly change the estimated differences between experimental conditions, which suggest that the response of infants with a larger number of trials did not influence our results substantially. Our sensitivity analyses supported the validity of our conclusions.

The number of trials included in an averaged measure influences the signal to noise ratio of the ERP and therefore the observed measure and its precision. Intra- and interindividual differences in an averaged ERP signal may simply reflect differences in the signal to noise ratio rather than differences in encoding and processing in response to stimuli.
The described relationship between the number of trials and the observed mean response is a potential methodological problem of all ERP studies with varying and small numbers of trials as the continuous decrease in mean observed amplitude after 10 trials suggests. For example, our model predicts a decrease of amplitude of 3.9 μV independent of experimental manipulation if the averaged mean is based on 14 instead of 10 trials. We recommend performing a variance component analysis and an assessment of the relationship between the number of trials and averaged response as a standard procedure at the beginning of the analysis of infant ERP data.

In addition, to control the influence of the number of trials on bias and precision of parameter estimates, we suggest to (1) weight a general linear mixed model analysis by the inverse of the variance associated with the number of trials used for the averaging if variance increases with increasing number of trials, (2) to include the number of trials as a linear and quadratic term in the analysis model, if a positive relationship between the number of trials and mean response can be observed and (3) to perform a sensitivity analysis by rerunning the model without weights to assess the influence of weights on parameter estimates. If sensitivity analyses lead to different conclusions the sources of discrepancies need to be assessed and discussed (e.g. does cognitive development influences the number of trials and could explain the discrepancies?).

General linear mixed models allow full case analyses and are less restricted on the assumptions of missingness: Even with missing data we will get unbiased parameter estimates if the probability of missing data on any variable is not related to its particular value and the pattern of missing data is predictable from other variables in the dataset (missing at random, [5]). Furthermore, GLMM allows including time(conditional) varying covariates and is flexible in modeling the covariance structure of the repeated observations [8].

As expected, we found in our study a relationship between the number of trials and both the estimates and the precision of the estimates in the Nc amplitude. Interestingly, the number of trials did not influence the observed averaged Nc latency: There was no relationship between the number of trials and the averaged response and only little influence on the variance of the responses. Furthermore, infants showed stronger differences in averaged latency across experimental conditions compared to averaged amplitudes. This suggests that in our study noise affects mainly the estimates of the observed amplitude but less the temporal location of the peak signal.

By using all infants in our sample we avoided a bias of our parameter estimates due to non-random dropouts. It is possible that the observed increase in latency in the angry compared to the neutral condition with direct gaze condition in the analysis with at least 10 trials only is due to a non-random sampling effect and therefore not representative for the population of 6 months old infants. For example, infants with 10 or more trials may be cognitively more developed and more attentive than the other infants. Missingness would not be at random (“Not missing at random” [5]) and generalizing the results from this subsample would be inaccurate.

Our approach enables the use of multifactorial designs in infant ERP studies, which are more likely to yield to unbiased estimates of the causal effects than an analysis based on a small proportion of the studied infants (cf. [27]). Our method would allow analyzing data sets of infant studies with smaller sample size than our 46 infants. It thus broadens the applicability of neuroscience methods in developmental research. We do not recommend using this method uncritically without analyzing the pattern of number of trials across experimental conditions, the relationship between number of trials and response and a careful assessment of the assumptions of the analysis model (e.g. by visual inspection of the residuals: [8]). Furthermore, our method is not a substitute for a good experimental design, which aims to collect a sufficient number of trials, and will not work if the signal to noise ratio is small.

The results may be biased if the experimental condition seriously influences the number of trials as we can assume that experimental condition will also correlate with ERP response. Conventional methods to analyze infants’ ERP data face similar risks of biased estimates because the number of trials still influences the ERP response beyond 10 trials. There was no evidence in our study but it would be desirable to further study the efficiency and consistency (the potential bias) of parameter estimates of ERP studies with varying numbers of trials in simulations comparing our method with traditional approaches using the data of individual trials of infant studies with sufficient number of trials per infant.

Acknowledgments

The research was supported by the Sofja Kovalevskaja Award granted by the Alexander von Humboldt Foundation, donated by the German Federal Ministry of Education and Research, to T. Striano. In addition, E. Parise was supported with a Humboldt Research Fellowship. We thank Christine Heinisch, Lena Schlossmacher, Angelika Becker and Anna Gröndahl for testing the infants. We are grateful to the infants and parents who participated and to the Universitätsfrauenklinik and the Klinikum St. Georg for assistance with recruitment.
References


