



Test-retest reliability of infant event related potentials evoked by faces

N.M. Munsters^{a,b,c,*}, H. van Ravenswaaij^d, C. van den Boomen^{a,b}, C. Kemner^{a,b,c}

^a Department of Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, The Netherlands

^b Department of Developmental Psychology, Utrecht University, Utrecht, The Netherlands

^c Department of Child and Adolescent Psychiatry, Brain Center Rudolf Magnus, University Medical Centre, Utrecht, The Netherlands

^d Educational Development & Training, Centre for Teaching and Learning, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Keywords:

Test-retest reliability
Faces
Event Related Potentials
Infants

ABSTRACT

Reliable measures are required to draw meaningful conclusions regarding developmental changes in longitudinal studies. Little is known, however, about the test-retest reliability of face-sensitive event related potentials (ERPs), a frequently used neural measure in infants. The aim of the current study is to investigate the test-retest reliability of ERPs typically evoked by faces in 9–10 month-old infants. The infants ($N=31$) were presented with neutral, fearful and happy faces that contained only the lower or higher spatial frequency information. They were tested twice within two weeks. The present results show that the test-retest reliability of the face-sensitive ERP components is moderate (P400 and Nc) to substantial (N290). However, there is low test-retest reliability for the effects of the specific experimental manipulations (i.e. emotion and spatial frequency) on the face-sensitive ERPs. To conclude, in infants the face-sensitive ERP components (i.e. N290, P400 and Nc) show adequate test-retest reliability, but not the effects of emotion and spatial frequency on these ERP components. We propose that further research focuses on investigating elements that might increase the test-retest reliability, as adequate test-retest reliability is necessary to draw meaningful conclusions on individual developmental trajectories of the face-sensitive ERPs in infants.

1. Introduction

Event related potentials (ERP) are often used to assess social, cognitive, and sensory information processing in infants. Previous ERP research has informed us on a group level about, for instance, the development of emotion discrimination (e.g. de Haan et al., 2004; Hoehl and Striano, 2010; Leppänen et al., 2007; Taylor-Colls and Pasco Fearon, 2015; Yrttiaho et al., 2014) and the influence of visual processing hereupon (van den Boomen et al., this issue; Vlamings et al., 2010a; Vlamings et al., 2010b). Recent studies are moving towards research on individual differences in social, cognitive, and sensory information processing in infants. For instance, currently there are large consortia that track behavioral and brain developmental trajectories (e.g. Consortium on Individual Development [CID]; European Autism Interventions – A Multicentre Study for Developing New Medications [EU-AIMS]). Research on individual differences might inform us on early neural markers (e.g. altered visual processing of emotional faces) of developmental disorders such as Autism Spectrum Disorder.

With the increase in studies on the individual development of the infant brain, it is crucial to test whether current brain measures are

methodologically suited for studying individual differences. One important methodological aspect is the stability of the ERP. The degree to which test scores of one individual are consistent over multiple measurements within a short period of time is assessed with test-retest reliability. Strong test-retest reliability is necessary to draw meaningful conclusions regarding developmental changes in longitudinal studies on brain development in infants. The test-retest reliability of visual ERPs in infants is to our knowledge however unknown. Therefore, the aim of the current study is to investigate the test-retest reliability of visual ERPs in infants. In line with previous research in our lab, we focus specifically on the test-retest reliability of visual ERPs evoked by emotional faces (filtered to contain specific visual information), as processing socially relevant information is crucial for early social and emotional development.

Surprisingly little is known about the test-retest reliability of visual ERPs. The few studies in children and adults that investigated the test-retest reliability of visual ERPs show mixed results (e.g. research in children and adults: Hämmerer et al., 2013; research in adults: Cassidy et al., 2012; Clayton and Larson, 2013; Huffmeijer et al., 2014; Nordin et al., 2011; Olvet and Hajcak, 2009; Segalowitz et al., 2010; van Deursen et al., 2009). Reliability of ERPs ranged from slight to almost

* Corresponding author at: Department of Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, The Netherlands.
E-mail address: n.m.munsters@gmail.com (N.M. Munsters).

perfect (Landis and Koch, 1977), with more variable and weaker test-retest reliability for latency than amplitude. Of these studies, there are two studies with adults that focused on ERPs evoked by faces (Cassidy et al., 2012; Huffmeijer et al., 2014). Both studies revealed that the reliability of the amplitude of the adult ERP evoked by emotional faces was almost perfect. Reliability of the latency was more variable and weaker, compared to the reliability of the amplitude of the face-sensitive ERP. In sum, whereas there is quite some variability in the test-retest reliability of ERPs, the reliability of ERPs evoked by faces in adults seems almost perfect for amplitude and more variable but generally fair to substantial for latency.

The characteristics of ERPs differ between infants and adults, with infants showing generally longer latencies and higher amplitudes, due to for example physiological differences such as myelination, folding, and the number of synapses. This even results in different waveforms related to face processing between infants and adults. The face-sensitive ERP component in adults (i.e. N170) has developmental precursors in infants: the N290 and P400. In addition, the negative central (Nc) is frequently associated with emotional face processing in infants (de Haan, 2007). Moreover, it is commonly recognized that noise levels differ strongly between infants and adults. Due to for instance increased movement, EEG data of infants contains more artifacts and higher noise levels (de Haan, 2007). As noise levels are crucial information for reliability calculations, conclusions on the test-retest reliability of ERPs in adults cannot be extended to infant ERP research.

The aim of the present study is to investigate in infants the test-retest reliability of the face-sensitive ERP components. We measured the ERPs evoked by fearful, happy, and neutral faces twice within two weeks. As the spatial frequency content of the faces is previously found to interact with the effects of emotion in children and adults (Pourtois et al., 2005; Vlamings et al., 2009, 2010a, 2010b), the faces were filtered to contain only the lower (representing the configuration of the face; global information) or higher (related to sharp edges in the face; local information) spatial frequency information. To give an overall view of the test-retest reliability of the face-sensitive ERP components in infants, we first performed test-retest reliability analyses for the amplitude of each component of interest (i.e. N290, P400, and Nc) in response to the faces (i.e. average over all emotional and spatial frequency conditions). Furthermore, because there is often interest in the effects of emotion and spatial frequency on the face-sensitive ERP components (van den Boomen et al., 2016; Vlamings et al., 2010a, 2010b), we also investigated the test-retest reliability of emotion and spatial frequency effects. We first investigated if there were main and interaction effects of emotion and spatial frequency. When such effects were present, we analyzed the test-retest reliability of these effects.

2. Methods

2.1. Participants

Seventy-seven 9–10 month-old infants were recruited from several communal registers in the Netherlands. The final sample consisted of 31 infants with sufficient data at both visits (18 males; at first visit: mean age is 299 days, range 279–317 days, SD 9 days; at second visit: mean age is 306 days, range 284–327 days, SD 9 days). An additional 46 infants were tested, but excluded from analyses due to refusal to wear the EEG cap, excessive motion, lack of attention, technical error, or medical reason. All infants were born full-term (> 37 weeks) and had no developmental delays or abnormalities in visual or auditory processing, as reported by the health-care system. The Medical Ethical Committee of the University Medical Centre of Utrecht approved the study protocol. The study is conducted in accordance with the Declaration of Helsinki. Parent(s) or guardian(s) of the infant gave written informed consent prior to participation. Children received a toy after participation.

2.2. Stimuli

Face stimuli consisted of photographs of 10 facial identities each depicted under 3 emotional conditions taken from the MacBrain Face Stimulus Set.¹ Face images included 5 males and 5 females, of which 6 European-American, 3 African-American and 1 Asian-American model. Face pictures were trimmed to remove external features (neck, ears, and hairline). Using Photoshop all stimuli were cropped, turned into grey-scale and matched for size (19.4×14.0 degrees of visual angle at a viewing distance of 57 cm). Faces had a fearful, neutral or happy facial expression and were filtered with a low- (LSF; < 2 cycles per degree; global) or high-pass (HSF; > 6 cycles per degree; local) spatial frequency filter. This created a 3 (emotion: fearful, neutral, and happy) \times 2 (spatial frequency: LSF and HSF) conditions design (Fig. 1).

2.3. Procedure

Infants were seated in a quiet and dimly lit room in a highchair positioned at eye level 57 cm from the computer screen while wearing the EEG cap. There were in total 60 stimuli (10 faces \times 3 emotion conditions \times 2 spatial frequency conditions). Stimuli were presented four times (divided over two blocks; in each block all conditions were presented an equal number of times), resulting in 240 trials. The stimuli were presented on a 23-in. screen with a resolution of 1920×1080 pixels, and a refresh rate of 60 Hz. Each trial consisted of a jittered inter-stimulus interval between 700 and 1000 ms followed by a face stimulus for 800 ms. A video camera was placed on top of the screen for online observation. When the infant was not looking at the screen, the experiment was paused and attention was reoriented by a sound played by the computer or a moving stimulus on the screen. Stimuli were presented until the infant became too fussy or bored to attend. Video recordings were additionally used for off-line coding of attention. Unattended trials (i.e. not looking with at least one eye to the stimulus, blinking and/or eyes not visible on the video during the first 500 ms of stimulus presentation) were discarded from analyses. The average number of attended trials was 175 for visit 1 (range: 130–215) and 154 for visit 2 (range: 100–236) for included participants. There were on average seven days between visits (range: 4–12 days).

2.4. Data analyses

2.4.1. ERP recording

EEG data was recorded with 32 electrodes (Active Two system, Biosemi) positioned at standard recording locations in a cap according to the international 10/20 system. During recording, EEG was sampled at a rate of 2048 Hz. Two extra electrodes, the CMS (common Mode Sense) and DRL (driven Right Leg), provided an active ground.

2.4.2. Preprocessing

Using Brain Vision Analyzer software (Brainproducts GmbH) and Matlab (The Mathworks, Natick, MA) we pre-processed the data. Data were resampled offline to 512 Hz, and filtered with a high-pass filter of .1 Hz (24 dB/oct), a low-pass filter of 30 Hz (24 dB/oct) and a notch filter of 50 Hz. In order to compute ERPs, epochs of 100 ms pre-stimulus (baseline) until 1000 ms post-stimulus were extracted from the continuous data. The data was demeaned, with baseline defined as 100 ms pre-stimulus until stimulus onset. Trials were removed in *single electrodes* when there were artifacts. Artifacts were defined as amplitudes below -200 or above $200 \mu\text{V}$; a difference of more than $200 \mu\text{V}$ within 100 ms; a difference of less than $3 \mu\text{V}$ within 200 ms; or a voltage

¹ Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for information concerning the stimulus set

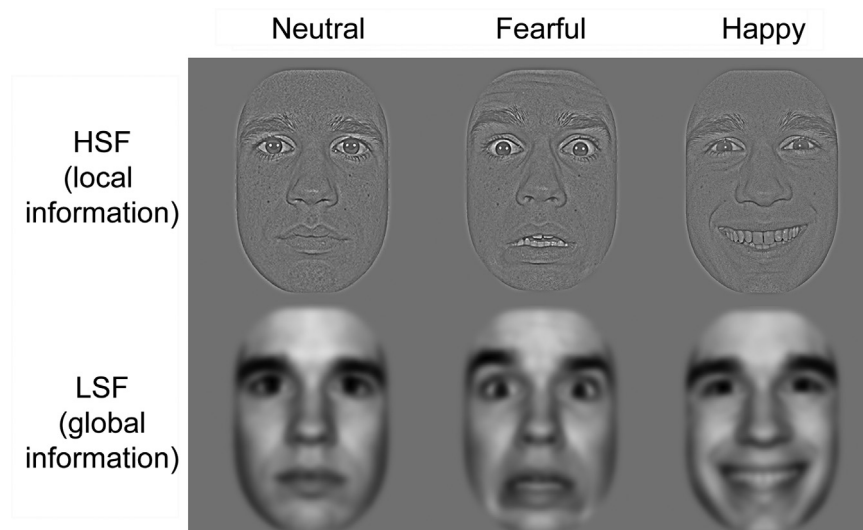


Fig. 1. Examples of the fearful, neutral, and happy low spatial frequency (LSF; global) and high spatial frequency (HSF; local) filtered face stimuli.

Table 1

Average included segments per condition.

	N290 and P400						Nc					
	Visit 1		Visit 2		Difference		Visit 1		Visit 2		Difference	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	Mean	SD	
Fear HSF	26	5	23	6	3	7	26	5	23	3	8	
Fear LSF	26	5	24	6	2	7	27	5	25	1	7	
Neutral HSF	28	5	25	6	3	8	29	5	26	3	8	
Neutral LSF	26	6	23	5	2	8	26	6	24	2	8	
Happy HSF	25	6	22	6	3	8	26	6	23	3	8	
Happy LSF	25	6	24	6	1	8	25	6	24	1	9	

change of more than 50 μV per sampling point. An electrode was rejected if there were less than 5 artifact-free trials. Trials were removed in all electrodes if the stimulus was unattended or contained an eye blink between 0 and 500 ms after onset (manually detected in the videos), or if more than 16% of electrodes contained artifacts as described above (based on previous research on face processing in infants, (e.g. Halit et al., 2003)). Finally, activity was referenced to the average of all included electrodes. For each stimulus condition an average of the ERP was created per electrode. Based on previous research in infants (Kobiella et al., 2007; Leppänen et al., 2007, 2009), participants were included in the data analyses if for each of the electrodes of interest (i.e. P3, PO3, O1, Oz, O2, PO4, P4, Cz, Fz, C3, C4, FC1, and FC2) there were at least 10 segments per condition included in the average. The average included segments per condition and electrode of interest was 26 for visit 1 and 24 for visit 2 (Table 1).

2.4.3. Component analyses

The components of interest were the N290, P400, and Nc. Mean activity within a time window of 200–325 ms (N290), 325–600 ms (P400), and 300–600 ms (Nc) was exported for further analyses on the amplitude of these components. Mean activity within a time window, instead of peak detection, was used because the N290, P400, and Nc did not have a clear peak in all infants. As a result, we could not determine peak latency and, as such, not investigate the test-retest reliability of peak latency. Electrodes of interest were based on previous research, resulting in the P3, PO3, O1, Oz, O2, PO4, and P4 for the N290 and P400. For the Nc, electrodes of interest were C3, Cz, C4, FC1, FC2, and Fz. Although based on previous research some additional electrodes of interest could be identified, those electrodes were excluded based on low data quality in most of the participants (i.e. P7, P8, T7, T8, F3, and

F4) or unclear components of interest (i.e. Pz). To limit the number of statistical comparisons, reliability analyses were performed for the average amplitude over all electrodes of interest. Data inspection showed that the electrode with the largest difference in amplitude between emotions differed across participants, thus the effects of emotion seem not limited to specific electrodes.

2.5. Statistical analyses

We investigated 1) the test-retest reliability of the face-sensitive ERP components overall and 2) the test-retest reliability of the effects of emotion and spatial frequency on the face-sensitive ERP components. To give an overall view of the test-retest reliability of the face-sensitive ERP components, we performed test-retest reliability analyses for the amplitude of each component of interest (i.e. N290, P400, and Nc) in response to the faces (i.e. average over all emotional and spatial frequency conditions). Before investigating the test-retest reliability of the effects of emotion and spatial frequency on the face-sensitive ERP components, we first investigated whether there were indeed group-effects of emotion and spatial frequency. Therefore, we performed repeated measures analyses of variance with visit (i.e. visit 1 and visit 2), emotion (i.e. happy, fearful, and neutral) and spatial frequency (i.e. LSF and HSF) as the independent variables and amplitude as the dependent variable. If main and/or interaction effects of emotion or spatial frequency were present, we performed paired sample *t*-tests and calculated difference scores between the conditions of these variables that were significantly different (e.g. happy versus fearful). Next, we analyzed the test-retest reliability of those difference scores, to provide insight in the test-retest reliability of the effects of emotion and spatial frequency on the face-sensitive ERP components. The test-retest

reliability of the overall and emotion and spatial frequency effects were analyzed in two steps: 1) intra-class-correlation coefficient and sigma-w and 2) correlations and *t*-tests.

2.5.1. Step 1: Intra-class-correlation coefficient and sigma-w

In the first step, the intra-class-correlation coefficient (ICC) and the sigma-w were calculated. The ICC is a relative measure of test-retest reliability, which describes the closeness of test scores from the same individual in two or more sessions within a short period of time, thus the consistency. We calculated for each variable of interest the ICC (using a two-way mixed model, absolute agreement, single measures), which is defined as the proportion of the total variance due to the between-subject variance. Because there is no consensus regarding reliability criteria (Weir, 2005), we quantified reliability as poor (ICC < .00), slight (ICC .00–.20), fair (ICC .21–.40), moderate (ICC .41–.60), substantial (ICC .61–.80), and almost perfect (ICC > .80) (Landis and Koch, 1977).

Sigma-w (σ_w) is the within-subject standard deviation and as such an absolute measure of reliability. We used this in addition to the ICC, to control for the between-subject variance. To calculate sigma-w, we first calculated for each subject the variance between visits for each variable. Next, the mean of the variable's variance is calculated and square rooted. This reflects the average change between visits. Poor test-retest reliability is reflected by a within-subject standard deviation (sigma-w) that is almost equal or larger than the between-subject standard deviation, as in this situation one participant could score relatively high on one visit and relatively low on another visit.

2.5.2. Step 2: Correlations and *t*-tests

In the second step, correlations and paired-sample *t*-tests between the first and second visit were performed. Correlations (Pearson's *r*) provide insight in whether the ranking of the amplitude between participants is stable between visits. We quantified the stability of ranking as very weak (*r* < .20), weak (*r* .20–.39), moderate (*r* .40–.59), strong (*r* .60–.79), and very strong (*r* > .79) (Evans, 1996). *T*-tests (i.e. the effect of visit) give an indication whether there is a systematic difference between the two visits. Since we repeated the same procedure twice and used the same equipment, we did not expect any group differences (i.e. systematic difference) between visit 1 and 2.

3. Results

3.1. Overall test-retest reliability of the N290, P400, and Nc

Results revealed that the test-retest reliability of the amplitude of the N290, P400, and Nc varied between moderate and substantial (Table 2; Figs. 2 and 3). The highest (i.e. substantial) test-retest reliability was found for the amplitude of the N290 (ICC = .76, *p* < .001). The within-subject standard deviation (σ_w = 4.1 μ V) was around half of the between-subject standard deviation (SD of 8 and 8.3 μ V). There was a strong correlation between visits (*r* = .77, *p* < .001) and no significant effect of visit (*t*(30) = −1.859, *p* = .073). Furthermore, there was moderate test-retest reliability of the P400 and Nc amplitude (P400: ICC = .56, *p* < .001; Nc: ICC = .57, *p* < .001). The within-subject standard deviation (P400: σ_w = 5.7 μ V; Nc: σ_w = 2.7 μ V) was around three-quarter of the between-subject standard deviation (P400: SD of 8.8 and 7.1 μ V; Nc: SD of 4.2 and 3.2 μ V). The results revealed a strong correlation between visits (P400: *r* = .69, *p* < .001; P400: *r* = .71, *p* < .001), but also a significant visit effect (P400: *t*(30) = 4.233, *p* < .001; Nc: *t*(30) = −4.714, *p* < .001).

3.2. Test-retest reliability of the effects of emotion and spatial frequency

Before investigating the test-retest reliability of the effects of emotion and spatial frequency on the face-sensitive ERP components, we first tested whether there were indeed emotion and/or spatial

frequency effects. The significant results from the repeated measures ANOVA's on N290, P400 and Nc amplitude are presented in Table 3 (all variables not included in the table were not significantly different: *p* > .05). We investigated the test-retest reliability of all significant effects (i.e. of the difference scores between the conditions that were significantly different).

Test-retest reliability of the emotion and the emotion * spatial frequency effects were all poor (ICC varied between −.33 and .15) (Table 2). The within-subject standard deviation (σ_w between 2.4 and 7.8 μ V) was almost similar to or higher than the within-subject standard deviation (SD between 2.4 and 7.4 μ V). There were no significant correlations between visits (*r* between −.31 and .16) and no significant effects of visit (*t*-test: all *p* > .05).

4. Discussion

The aim of the present study was to investigate the test-retest reliability of the face-sensitive ERP components in infants. We measured ERPs in response to emotional (i.e. neutral, fearful and happy) faces that were filtered for specific visual information (i.e. lower or higher spatial frequency information), twice within two weeks. First, the test-retest reliability was analyzed for the amplitude of each component of interest (i.e. N290, P400 and Nc) in response to the faces overall. Secondly, we investigated the test-retest reliability of the main and interaction effects of emotion and spatial frequency (i.e. difference scores between conditions that showed a main and/or interaction effect).

The present results show that the overall test-retest reliability of the face-sensitive ERP components in infants is substantial for the amplitude of the N290. For the amplitude of the P400 and Nc moderate test-retest reliability was found. The results of the amplitude of the face-sensitive ERP components indicate lower test-retest reliability in infants than in adults, as previous research showed almost perfect test-retest reliability for adults' face-sensitive ERPs (Cassidy et al., 2012; Huffmeijer et al., 2014). These findings are in line with previous findings on the N200, which showed that – although there was no significant difference between age groups in test-retest reliability – children showed fair to moderate test-retest reliability whereas reliability was moderate to substantial for adults (Hämmerer et al., 2013).

The lower test-retest reliability in infants compared to adults could relate to the age differences in the face-sensitive ERPs. Probably most important, infant EEG data contains more artifacts and higher noise levels (de Haan, 2007). This leads to a higher within-subject variance, which negatively affects the intra-class correlations. Another explanation for the lower test-retest reliability could be the number of trials. The number of trials is shown to increase the test-retest reliability of ERPs (i.e. VPP and P3) in previous research in adults (Huffmeijer et al., 2014). There are two studies on the test-retest reliability of face-sensitive ERPs in adults (Cassidy et al., 2012; Huffmeijer et al., 2014), which both included a higher number of trials per participant (i.e. on average approximately 30 to 50 artifact-free trials per condition) than the current standard in infant research (a minimum of 10 trials per condition) used in the present research. The current standard for the minimum number of trials in infant research is lower than in adults, because of the behavioral tendencies (e.g. short attention span, frequent movements) of infants. Using a higher number of trials in infant research might increase the test-retest reliability. The number of trials could also affect ERP amplitude: in a previous study, the amplitude of the Nc decreased when more trials were included in the average (Hoehl and Wahl, 2012). As a consequence, differences in the number of trials between visits could result in lower test-retest reliability. On the whole, using a varying number of trials within and between infants, the present research provides an indication of the test-retest reliability of infants face-sensitive ERPs as measured according to frequently used ERP methods in infant research.

Although the test-retest reliability of the face-sensitive ERP compo-

Table 2

Test-retest reliability results of the (effects of emotion and spatial frequency on the) face-sensitive ERP components.

		Visit 1 Mean (SD)	Visit 2 Mean (SD)	ICC	Sigma-W (σ_w)	Correlation (r)	T-test (t)
Overall							
N290	Average of all conditions	11.6 (8.3)	9.7 (8.0)	.76***	4.1	.77***	1.859
P400	Average of all conditions	22.0 (8.8)	17.1 (7.1)	.58***	5.7	.69***	4.233***
Nc	Average of all conditions	−9.9 (4.2)	−7.5 (3.2)	.57***	2.7	.71***	−4.714***
Emotion effects							
N290	Happy - Neutral	3.8 (4.4)	3.1 (3.3)	−.15	4.1	−.15	.615
	Happy - Fear	2.3 (3.9)	.8 (3.6)	.08	3.7	.08	1.632
	Fear - Neutral	1.5 (5.1)	2.4 (4.3)	−.13	5.0	−.13	−.669
P400	Happy - Neutral	1.6 (3.7)	2.6 (4.5)	.03	4.1	.03	−.971
	Happy - Fear	.9 (4.0)	2.0 (4.4)	−.09	4.4	−.09	−.961
Nc	Happy - Neutral	−1.0 (2.4)	−.8 (2.4)	−.04	2.4	−.04	−.337
Emotion * spatial frequency effects							
N290	HSF: Happy - Neutral	5.2 (7.3)	6.0 (4.4)	.05	5.8	.05	−.515
	HSF: Happy - Fear	2.9 (6.2)	1.8 (5.3)	−.08	6.0	−.08	.691
	HSF: Fear - Neutral	2.4 (7.4)	4.2 (5.9)	−.13	7.1	−.14	−1.010
P400	HSF: Happy - Neutral	3.3 (5.3)	5.5 (5.1)	.15	4.9	.16	−1.756
	HSF: Happy - Fear	1.3 (6.3)	3.1 (5.4)	.03	5.8	.03	−1.244
	HSF: Fear - Neutral	2.1 (7.2)	2.4 (6.6)	−.33	7.8	−.31	−.152

Note. HSF is higher spatial frequency filter; Amplitude (mean, SD and sigma-W) is given in μV .*** $p < .001$;

nents in infants is lower than in adults, it is moderate to substantial. However, the test-retest reliability of the effects of emotion and spatial frequency on the face-sensitive ERP components in infants is poor. The poor test-retest reliability could relate to the modest effects of emotion and/or spatial frequency effects on the face-sensitive ERPs. There is a higher signal to noise ratio needed to reliably measure those relatively small effects. There is to our knowledge no previous research on the test-retest reliability of emotion and/or spatial frequency effects on the face-sensitive ERP components. The current results are not surprising when looking at the research at a group level on emotion discrimination in infancy this far. That is, the previous results on the ability to discriminate emotions in infancy are somewhat inconsistent (e.g. de Haan et al., 2004; Hoehl and Striano, 2010; Leppänen et al., 2007; Taylor-Colls and Pasco Fearon, 2015; Yrttiaho et al., 2014). These mixed findings could partly relate to differences in methods between studies, such as differences in the presented emotions. Another explanation could be that studies often use a small sample size and low number of trials. This, especially together with low test-retest reliability, could result in the sample not representing the population very well. Nonetheless, in the present study, there were no significant interaction effects of visit with emotion and/or spatial frequency, which indicates that the emotion and spatial frequency effects are replicable at the group level.

Important implications for further research can be drawn from the

present results. According to the Landis and Koch (1977) criteria of test-retest reliability, the amplitude of the face-sensitive ERP components has moderate to substantial reliability in infants compared to almost perfect reliability in adults. Yet, there is no consensus on reliability criteria and one could doubt if almost perfect test-retest reliability is achievable in infants. To be able to draw meaningful conclusions on the individual developmental trajectories, we would advise focusing on the variables with at least moderate test-retest reliability, such as the amplitude of the N290, P400 and Nc. As none of the difference scores in the present study showed adequate test-retest reliability, we would suggest that further research focuses on elements that might increase the test-retest reliability of infants' face-sensitive ERPs. A starting point for further research could be the influence of the number of trials on the test-retest reliability of the face-sensitive ERP components in infants. It could however be that increasing the number of trials leads to problems with feasibility, because of the behavioral tendencies (e.g. short attention span, frequent movements, falling asleep or crying) of infants. Larger number of trials might be possible, given that we had enough trials to compare the ERP responses between six conditions in the present study. When fewer conditions are investigated the number of trials per condition can be increased. Furthermore, Hoehl and Wahl (2012) provide several suggestions that might increase the number of trials with good data quality.

There are some limitations to the present study. Firstly, we used

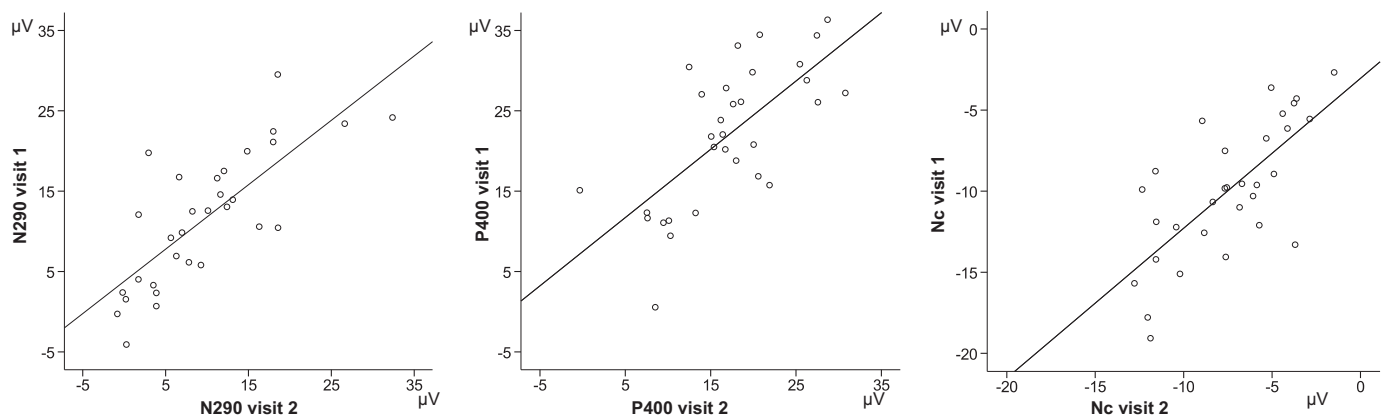


Fig. 2. Correlation between visit 1 and 2 for the N290, P400, and Nc.

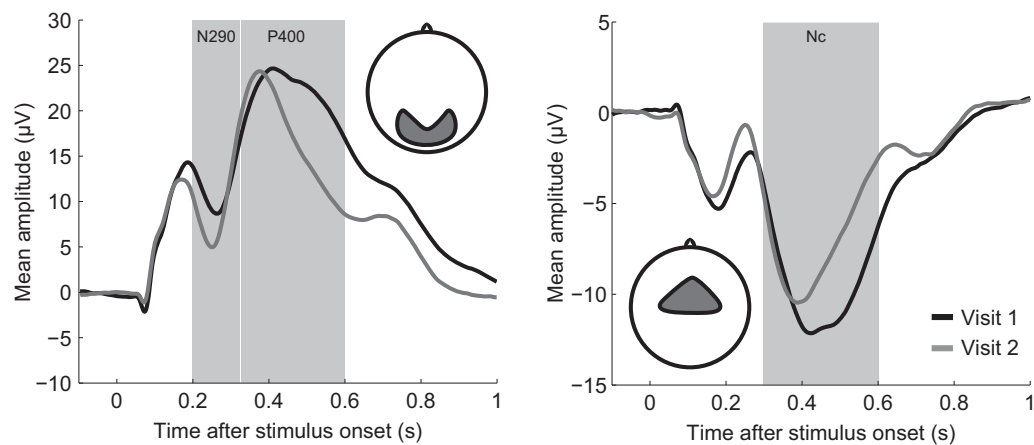


Fig. 3. N290, P400, and Nc on visit 1 and 2.

Table 3
Significant results from the Repeated Measures ANOVA.

	F	Effect-size	Comparison between conditions
Visit effects			
P400	17.917***	.374	visit 1 > visit 2***
Nc	22.220***	.426	visit 1 > visit 2***
Emotion effects			
N290	23.520***	.439	N > H***; F > H*; N > F**
P400	8.147**	.214	N < H***; F < H**
Nc	3.259*	.098	N < H*
Emotion * spatial frequency effects			
N290	6.956**	.188	HSF: N > H***; F > H**; N > F***
P400	9.680***	.244	HSF: N < H***; F < H**; N < F**

Note. all other comparisons were non-significant, > indicates a larger amplitude; N is neutral, F is fearful, H is happy; HSF is higher spatial frequency
* p < .05,
** p < .01,
*** p < .001,

faces that were filtered for specific visual (i.e. spatial frequency) content. It could be that the test-retest reliability is lower for filtered faces than for non-filtered faces, as filtered faces contain less visual information. Further research should investigate whether the test-retest reliability is different for non-filtered faces. Secondly, although the infants were tested twice with the same procedure and equipment, there is still some variance between the visits (i.e. systematic difference) for the overall ERP responses. The smaller P400 and Nc amplitude on the second visit could be the result of development, but also of habituation and/or a learning effect taking place between the visits. As a result of these systematic differences, the test-retest reliability of the P400 and Nc is moderate, while there is a strong correlation. There was no significant difference between visits for the effects of emotion and spatial frequency on the ERP responses. Nonetheless, as the manipulations of emotion and spatial frequency have modest effects on the face-sensitive ERP components, the difference scores between the emotion and/or spatial frequency conditions were possibly too small and the between subject-variance too high to detect differences between visits. Therefore, we cannot firmly conclude there was no systematic difference between visits for the effects of emotion and spatial frequency on the ERP responses. A third limitation of the present study is the variance in the time period between visits (i.e. 4–12 days), as the time period between visits might affect the test-retest reliability. For most children (N = 24) there were seven days between the visits, therefore we could not investigate the role of the time period between visits on the results. When only investigating the test-retest reliability for the children retested after seven days, the test-retest reliability of the

N290, P400, and Nc increases slightly (ICC increase between .04 and .10). The test-retest results for the effects of emotion and spatial frequency stayed similar, and still no systematic differences between visits were found. Thus, in the present study there is little evidence for an influence of days between visits on the test-retest reliability.
To conclude, the N290, P400, and Nc amplitude have moderate to substantial test-retest reliability in infants. However, even though emotion and spatial frequency effects on these ERP components are replicable at the group level, none of these effects show adequate test-retest reliability in infants. Before investigating individual developmental trajectories of the face-sensitive ERP components from infancy onwards, more research is needed to validate the present results and to investigate elements (e.g. the number of trials) that might increase the test-retest reliability. If we can determine such elements, we might be able to adjust our methods in a way that would result in adequate test-retest reliability of emotion and spatial frequency effects on the face-sensitive ERP components in infants. This is necessary to draw meaningful conclusions on individual developmental trajectories of emotion discrimination in longitudinal research.

Acknowledgements

The study was financed through the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Grant Number 024.001.003 awarded to author CK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank the parents and infants for participating and all employees at the ChildResearchCenter of Utrecht University for help with data collection.

References

Cassidy, S.M., Robertson, I.H., O'Connell, R.G., 2012. Retest reliability of event-related potentials: evidence from a variety of paradigms. *Psychophysiology* 49 (5), 659–664. <http://dx.doi.org/10.1111/j.1469-8986.2011.01349.x>.
Clayson, P.E., Larson, M.J., 2013. Psychometric properties of conflict monitoring and conflict adaptation indices: response time and conflict N2 event-related potentials. *Psychophysiology* 50 (12), 1209–1219. <http://dx.doi.org/10.1111/psyp.12138>.
de Haan, M., 2007. *Infant EEG and Event-Related Potentials*. Psychology Press.
de Haan, M., Belsky, J., Reid, V., Volein, A., Johnson, M.H., 2004. Maternal personality and infants' neural and visual responsivity to facial expressions of emotion. *J. Child Psychol. Psychiatry, Allied Discip.* 45 (7), 1209–1218. <http://dx.doi.org/10.1111/j.1469-7610.2004.00320.x>.
Evans, J.D., 1996. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole.
Halit, H., De Haan, M., Johnson, M.H., 2003. Cortical specialisation for face processing: face-sensitive event-related potential components in 3- and 12-month-old infants. *NeuroImage* 19 (3), 1180–1193.
Hämmerer, D., Li, S.-C., Völkle, M., Müller, V., Lindenberger, U., 2013. A lifespan

- comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. *Psychophysiology* 50 (1), 111–123. <http://dx.doi.org/10.1111/j.1469-8986.2012.01476.x>.
- Hoehl, S., Striano, T., 2010. The development of emotional face and eye gaze processing. *Dev. Sci.* 13 (6), 813–825. <http://dx.doi.org/10.1111/j.1467-7687.2009.00944.x>.
- Hoehl, S., Wahl, S., 2012. Recording infant ERP data for cognitive research. *Dev. Neuropsychol.* 37 (3), 187–209. <http://dx.doi.org/10.1080/87565641.2011.627958>.
- Huffmeijer, R., Bakermans-Kranenburg, M.J., Alink, L.R.A., van IJzendoorn, M.H., 2014. Reliability of event-related potentials: the influence of number of trials and electrodes. *Physiol. Behav.* 130, 13–22. <http://dx.doi.org/10.1016/j.physbeh.2014.03.008>.
- Kobiella, A., Grossmann, T., Reid, V.M., Striano, T., 2007. The discrimination of angry and fearful facial expressions in 7-month-old infants: an event-related potential study. *Cogn. Emot.* 22 (1), 134–146. <http://dx.doi.org/10.1080/02699930701394256>.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1), 159–174.
- Leppänen, J.M., Moulson, M.C., Vogel-Farley, V.K., Nelson, C.A., 2007. An ERP study of emotional face processing in the adult and infant brain. *Child Dev.* 78 (1), 232–245. <http://dx.doi.org/10.1111/j.1467-8624.2007.00994.x>.
- Leppänen, J.M., Richmond, J., Vogel-Farley, V.K., Moulson, M.C., Nelson, C.A., 2009. Categorical representation of facial expressions in the infant brain. *Infancy* 14 (3), 346–362. <http://dx.doi.org/10.1080/15250000902839393>.
- Nordin, S., Andersson, L., Olofsson, J.K., McCormack, M., Polich, J., 2011. Evaluation of auditory, visual and olfactory event-related potentials for comparing interspersed- and single-stimulus paradigms. *Int. J. Psychophysiol.: Off. J. Int. Organ. Psychophysiol.* 81 (3), 252–262. <http://dx.doi.org/10.1016/j.ijpsycho.2011.06.020>.
- Olvet, D.M., Hajcak, G., 2009. Reliability of error-related brain activity. *Brain Res.* 1284, 89–99. <http://dx.doi.org/10.1016/j.brainres.2009.05.079>.
- Pourtois, G., Dan, E.S., Grandjean, D., Sander, D., Vuilleumier, P., 2005. Enhanced extrastriate visual response to bandpass spatial frequency filtered fearful faces: time course and topographic evoked-potentials mapping. *Human. Brain Mapp.* 26 (1), 65–79. <http://dx.doi.org/10.1002/hbm.20130>.
- Segalowitz, S.J., Santesso, D.L., Murphy, T.I., Homan, D., Chantzantonou, D.K., Khan, S., 2010. Retest reliability of medial frontal negativities during performance monitoring. *Psychophysiology* 47 (2), 260–270. <http://dx.doi.org/10.1111/j.1469-8986.2009.00942.x>.
- Taylor-Colls, S., Pasco Fearon, R.M., 2015. The effects of parental behavior on infants' neural processing of emotion expressions. *Child Dev.* 86 (3), 877–888. <http://dx.doi.org/10.1111/cdev.12348>.
- van den Boomen, C., Munsters, N.M., Kemner, C., this issue. Emotion processing in the infant brain: the importance of local information. Manuscript submitted for publication.
- van Deursen, J.A., Vuurman, E.F.P.M., Smits, L.L., Verhey, F.R.J., Riedel, W.J., 2009. Response speed, contingent negative variation and P300 in Alzheimer's disease and MCI. *Brain Cogn.* 69 (3), 592–599. <http://dx.doi.org/10.1016/j.bandc.2008.12.007>.
- Vlamings, P.H.J.M., Goffaux, V., Kemner, C., 2009. Is the early modulation of brain activity by fearful facial expressions primarily mediated by coarse low spatial frequency information? *J. Vision.* 9 (5), 12.1–13. <http://dx.doi.org/10.1167/9.5.12>.
- Vlamings, P.H.J.M., Jonkman, L.M., Kemner, C., 2010a. An eye for detail: an event-related potential study of the rapid processing of fearful facial expressions in children. *Child Dev.* 81 (4), 1304–1319. <http://dx.doi.org/10.1111/j.1467-8624.2010.01470.x>.
- Vlamings, P.H.J.M., Jonkman, L.M., van Daalen, E., van der Gaag, R.J., Kemner, C., 2010b. Basic abnormalities in visual processing affect face processing at an early age in autism spectrum disorder. *Biol. Psychiatry* 68 (12), 1107–1113. <http://dx.doi.org/10.1016/j.biopsych.2010.06.024>.
- Weir, J.P., 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res.* 19 (1), 231–240. <http://doi.org/10.1519/15184.1>.
- Yrttiaho, S., Forssman, L., Kaatila, J., Leppänen, J.M., 2014. Developmental precursors of social brain networks: the emergence of attentional and cortical sensitivity to facial expressions in 5 to 7 months old infants. *PLoS ONE* 9 (6), e100811. <http://dx.doi.org/10.1371/journal.pone.0100811>.