HHS Public Access

Author manuscript

Dev Neuropsychol. Author manuscript; available in PMC 2021 October 20.

Published in final edited form as:

Dev Neuropsychol. 2020 September ; 45(6): 341–366. doi:10.1080/87565641.2020.1833208.

Test-Retest Reliability of Electroencephalographic Measures of Performance Monitoring in Children and Adults

Mei-Heng Lin[a], **Patricia L Davies**[a,b], **Jaclyn Stephens**[a,b], **William J Gavin**[b]

[a]Department of Occupational Therapy, 1573 Campus Delivery, Colorado State University, Fort Collins, CO 80523, USA

[b]Department of Molecular, Cellular & Integrative Neuroscience, 1680 Campus Delivery, Colorado State University, Fort Collins, CO 80523, USA

Abstract

This study examined the test-retest reliability of the error-related negativity (ERN) and error positivity (Pe) amplitudes using a Flanker task in 118 neurotypical children and 53 adults before and after latency jitter adjustments. The reliability of the ERN and Pe amplitudes was moderate for children and moderate to strong for adults. The latency variability adjustment did not improve the reliability of the ERN and Pe amplitudes for either group, suggesting that latency variability may be a trait-like measure. For comparison purposes, the reliability of the stimulus-locked ERPs was strong for correct trials, yet the reliability was weak for incorrect trials.

Keywords

error-related negativity (ERN); test-retest reliability; trial-to-trial variability; performance monitoring

Introduction

Performance monitoring is a set of mental processes including the evaluation of ongoing behavior, detection of performance errors, and initiation of post-error behavioral adjustment (Coles, Scheffers, & Holroyd, 2001). Collectively, these processes allow individuals to perform goal-directed behaviors. Electroencephalography (EEG) has been used to understand underlying neural mechanisms of performance monitoring, which is indicated by two event-related potential (ERP) components, namely error-related negativity (ERN), and error positivity (Pe). The ERN component is a frontally distributed negative voltage deflection which usually peaks between 0–80 ms for adults and between −30 – 120 ms for children following incorrect responses. It has been associated with error detection, conflict monitoring, motivational significance of errors, or emotional response to errors (Coles et al.,

Corresponding author: Patricia L Davies, Department of Occupational Therapy, 1573 Campus Delivery, Colorado State University, Fort Collins, CO 80523, USA, patricia.davies@colostate.edu, Phone: 970-491-7294.

Data Statement: The data will be made available to interested researchers. To access the data the researchers should directly contact the corresponding author – patricia.davies@colostate.edu.

Conflict of Interest: There are no known conflicts of interest.

2001; Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Larson, Clayson, & Clawson, 2014; Olvet & Hajcak, 2008; Weinberg, Dieterich, & Riesel, 2015; Yeung, Botvinick, & Cohen, 2004). Although the ERN has been shown to reflect activity of neural systems responsible for error awareness (Scheffers & Coles, 2000), whether the ERN is directly related to conscious awareness of errors remains controversial (Nieuwenhuis et al., 2001; Wessel, 2012). Studies from functional magnetic resonance imaging (*f*MRI) and EEG dipole modeling suggest that the primary neural generator of the ERN is located at anterior cingulate cortex (ACC; Carter et al., 1998; Coles et al., 2001; Holroyd, Dien, & Coles, 1998; Mathalon, Whitfield, & Ford, 2003; van Veen & Carter, 2002). The Pe component is a slow positive deflection that follows the ERN and peaks at 300–500ms following incorrect responses (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991). The Pe has been associated with conscious cognitive processing of errors, error awareness, and initiation of post-error adjustment (Overbeek et al., 2005; van Veen & Carter, 2006; Davies, Segalowitz, Dywan, & Pailing, 2001; Ridderinkhof, Ramautar, & Wijnen, 2009; Falkenstein et al., 2000; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001), and its primary neural generator is believed to be the rostral ACC (Herrmann, Rommler, Ehlis, Heidrich, & Fallgatter, 2004). Studies have shown a significant relationship between the Pe amplitude and the post-error slowing (i.e. a prolonged response time) following errors to ensure the overall performance accuracy (Nieuwenhuis et al., 2001; Overbeek, Nieuwenhuis, & Ridderinkhof, 2005).

Several studies have demonstrated that individuals with neurological disorders show atypical ERN and Pe amplitudes compared to neurotypical peers. For instance, the ERN amplitude has been shown to be smaller in individuals with schizophrenia (Bates, Liddle, Kiehl, & Ngan, 2004; Kim et al., 2006; Morris, Heerey, Gold, & Holroyd, 2008; Morris, Yee, & Nuechterlein, 2006), traumatic brain injury (Larson, Kaufman, Kellison, Schmalfuss, & Perlstein, 2009), and depression (Ruchsow et al., 2006). Likewise, the Pe amplitude has been shown to be smaller in individuals with attention-deficit hyperactivity disorders (Van De Voorde, Roeyers, & Wiersema, 2010) and schizophrenia (Rabella et al., 2016). In contrast, a larger ERN has been reported in individuals with obsessive compulsive disorders (Carrasco et al., 2013) and anxiety disorders (Ladouceur, Dahl, Birmaher, Axelson, & Ryan, 2006). These findings suggest that the ERN and Pe are trait-like measures and may serve as biomarkers for screening individuals with neurological disorder or psychiatric conditions. As a result, there is a growing body of literature investigating the psychometric properties of the ERN and Pe components (Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013; Foti, Kotov, & Hajcak, 2013; Meyer, 2017).

Several studies have investigated the test-retest reliability of the ERN and Pe components in neurotypical adults and reported strong test-retest reliability of the ERN and Pe amplitudes across two sessions with intervals between sessions ranging from 20 minutes to 2 years. Segalowitz et al. (2010) in study 1 showed moderate to strong ERN test-retest reliability (r = 0.76, $p < .05$; intraclass correlation coefficient (ICC) = 0.73, $p < .01$) in neurotypical adults with 20 minutes test-retest interval within one session, avoiding measurement error due to reapplying the cap. In study 2, twenty-eight 15-year-old adolescent boys participated in a test-retest with a 3 to 6 week interval between session resulting in a moderate ERN reliability (r = 0.63, $p < .01$; ICC = 0.59, $p < .01$ (Segalowitz et al., 2010) . Olvet & Hajcak (2009a) examined the reliability on the ERN amplitude collected from two visits with two

weeks apart using the flanker task on 45 undergraduate students. The ERN amplitude (peak measure) demonstrated strong test-retest reliability ($r = 0.74$, $p < .001$; ICC = 0.70, $p < .001$). Similarly, strong test-retest reliability was observed in the Pe amplitude (area measure; $r = 0.75$, $p < .001$; ICC = 0.75, $p < .001$). Moreover, Cassidy et al. (2012) reported that the ERN and Pe amplitudes (peak measure) collected from two separate visits with a month between sessions demonstrated strong test-retest reliability (ERN: $r = 0.75$, $p < .001$; ICC = 0.74, $p < .001$; Pe: $r = 0.74$, $p < .001$; ICC = 0.71, $p < .001$) on 25 neurotypical adults using the flanker task. Similarly, Weinberg et al. (2011) also demonstrated moderate to strong test-retest reliability of the ERN on two sessions separated 1.5 to 2 years in 26 undergraduate students ($r = 0.65$, $p < .01$; ICC = 0.62, $p < .01$). Despite consistent findings in adult literature, little research exists examining the test-retest reliability of the ERN and Pe components in children. We only found one study conducted by Meyer et al. (2014), and their findings demonstrated that the ERN had moderate to strong test-retest reliability in 44 children aged 8 to 13 years with testing completed 2 years apart ($r = 0.63$, $p < .01$). However, the ERN amplitude has been shown as a developmental phenomenon, such that the amplitude gradually increased across the age range from 7 to 18 years old (Davies et al., 2004; Gavin, Lin, & Davies, 2019). Therefore, the changes in the ERN amplitude across ages might confound the reliability coefficient reported in the Meyer et al. (2014).

Moreover, when investigating the psychometric properties of the ERPs in children and adults, it is important for researchers to consider other sources of variance in order to obtain robust measure of underlying cognitive processes (Segalowitz & Dywan, 2009). Gavin and Davies (2008) proposed a model to conceptualize five potential sources of variance that contribute to any psychophysiological measures (PM) such as ERPs, the model is presented as:

$$\text{PM} = \text{Effect}_{\text{STIMULUS}} + \text{Effect}_{\text{STATE}} + \text{Effect}_{\text{TRAIT}} + \text{Effect}_{\text{PM\_PROCESSING}} + \text{Measurement error}$$

Specifically, these variables are: (1) $\text{Effect}_{\text{STIMULUS}}$, the influence of the stimuli being presented (e.g. paradigm researchers used for testing); (2) $\text{Effect}_{\text{STATE}}$, the state of individuals at the time of testing (e.g. fatigue); (3) $\text{Effect}_{\text{TRAIT}}$, the trait(s) of individuals (e.g. age, sex, or cognitive capacities being investigated); (4) $\text{Effect}_{\text{PM\_PROCESSING}}$, the signal processing parameters implemented to obtain the ERPs; and (5) measurement error, any unaccounted variance. When examining the test-retest reliability of ERPs, such as the ERN and Pe amplitudes, researchers strive to control for - or minimize - the variance associated with stimuli, state, trait, and data processing parameters. For instance, researchers may utilize the same testing paradigm, make sure participants were emotionally stable and physically comfortable during the time of testing, set age and sex as covariates, and standardize the signal processes procedure across two sessions.

Another source of variance related to processing data is the number of trials used to produce averaged ERN amplitude and can play a critical role when investigating reliability. For example, Olvet & Hajcak (2009b) and Pontifex et al. (2010) suggested 6 trials are needed to produce stable split-half reliability of ERN amplitude in neurotypical groups. Larson, Baldwin, Good, & Fair (2010) demonstrated that more than 14 trials are required to obtain

adequate test–retest reliability in adults whereas Baldwin, Larson, & Clayson (2015) suggested that at least 13 trials are required for clinical population.

However, when examining the test-retest reliability in adults and children one source of variance embedded in the traditional ERP data analyses which most researchers do not control for is the trial-to-trial variation in latency (i.e. latency jitter). The latency jitter, could in turn, be regarded as a source of unaccounted variance (measurement error), and confound the results. Specifically, by using the traditional data analysis approach, researchers assume that the ERP evoked by certain events (e.g. incorrect button presses) are invariant and time-locked over multiple event presentations. Thus, averaging ERPs across multiple segments reduces irrelevant background noise and retains the brain responses evoked by the events. By making this assumption, researchers overlook the impact of the trial-to-trial latency jitter on the averaged ERP amplitude (DuPuis et al., 2014; Luck, 2014). Particularly, the considerable amount of latency jitter across segments can attenuate the amplitude of averaged ERP for a single individual (Luck, 2014; Lukie, Montazer-Hojat, & Holroyd, 2014; Unsal & Segalowitz, 1995; van Boxtel, 1998). Additionally, latency jitter has been shown to be larger in children compared to adults (Gavin et al., 2019; Lukie et al., 2014; Segalowitz & Dywan, 2009). For example, Lukie et al. (2014) explored the developmental changes of an ERP component related to decision-making, namely the reward positivity, in children (8-13 years), adolescents (14-17 years), and young adults (18-23 years). In this article, the researchers' visual inspection of the averaged ERP of children revealed greater latency variability compared to adolescents and adults. Correction of the latency variability was accomplished by re-aligning the times of the reward positivity of the averaged ERPs across individuals of each age group to create new grand averages. While the new grand-averaged ERP figures illustrated that latency jitter attenuated the grand-average ERP amplitude, particularly in children, this approach to the latency jitter correction did not alter the underlying statistical results (Lukie et al., 2014).

In fact, two recent studies demonstrated that correcting the averaged ERP by accounting for the trial-to-trial variation in latency of the ERP component does reduce noise and measurement error (Tabachnick, et al., 2018 and Gavin et al., 2019). Tabachnick and colleagues examined the relationship between the ERN and depressive symptoms among children involved in the Child Protective Services and a group of control children, all around 8 years of age. They used Woody-filter procedures similar to those described by Woody (1967) and based on visual inspection determined that the Woody-filter procedures reduce the latency variability in the grand-average ERN amplitude in their sample without altering the statistical group differences. In the Gavin, et al. study, an Adaptive Woody filter technique was used to adjust for latency variability in the ERN data collected from 240 participants aged 7 to 25 years. Implementation of the Adaptive Woody filter not only allowed for the measurement of variability, but also removed the latency variability effect from each individual's averaged ERP. This allowed for the assessment of developmental trends of the ERN without the confound of trial-to-trial variability. Three principle findings were discussed in this study. First, using several measures of fit and intra-individual variability, the presence of trial-to-trial latency variability of the ERN component was found in individuals at all ages, including adults, but was greatest in young children. Second, for each latency variability measure, the degree of trial-to-trial variability was shown to decrease

as the age of the participants increased. Lastly, the success of the Adaptive Woody filter technique in removing the trial-to-trial variability was demonstrated in a straightforward manner in the changes in the measures of fit and intra-individual variability before and after applying the filter. These findings suggest that applying an Adaptive Woody filter technique may also improve on the temporal reliability of the ERN and Pe measures.

Therefore, the purpose of the present study was to examine the test-retest reliability of the ERN and Pe amplitude before and after correcting for the trial-to-trial latency variability in neurotypical children and adults. Specifically, we utilized a speeded, forced-choice visual flanker task to elicit errors for each participant who completed 2 sessions, 1-3 weeks apart. This study focuses on three research questions. The first question asks: Can we replicate developmental differences in the ERN and Pe amplitude between adults and children using a speeded version of the flanker task? In keeping with Gavin, et al., 2019, despite a change in the flanker task procedures, we expect that adults will again demonstrate larger ERN and Pe amplitudes than children before and after latency jitter correction. The second research question asks: What is the test-retest reliability of the ERN and Pe amplitude in children and adults? Based on the literature review, we hypothesize that the test-retest reliability of the ERN and Pe amplitude will be stronger in adults compared to children. Our third research question asks: Does the implementation of latency jitter correction via the implementation of Woody Filter technique improve the reliability of the ERN and Pe amplitudes? We hypothesize that the reliability will be stronger after the latency jitter correction, due to correcting for the variations in the ERN and Pe components at single trial level. Additionally, we investigated the test-retest reliability of the mid-to-late ERP components elicited by the stimulus to evaluate the consistency of overall attention and adherence to the task across sessions. This was done as a procedural control for evaluating the validity of the reliability of the ERN and Pe.

Methods

Participants

A total of 241 participants - 74 neurotypical adults, aged 18 to 30 years, and 167 typically-developing children, aged 8 to12 years, - were recruited from the university and local community through campus emails, flyers, research subject pool of the Psychology department, and word of mouth. All participants were screened for neurological disorders and use of psychopharmaceutical drugs (e.g., antidepressants) by parent- or self- report. Application of exclusion criteria resulted in a few participants being excluded from data analysis; 3 adults and 12 children due to parent- or self-reported diagnoses of brain injury, learning disability, reading disability, depression, or attention-deficit hyperactivity disorders. Additionally, 1 adult and 9 children were excluded due to failure to complete one or both sessions. Furthermore, to keep the number of trials within the reasonable range (6 - 14 trials) suggested by previous studies (Larson et al., 2010; Baldwin et al., 2015; Olvet & Hajcak, 2009b; Pontifex et al., 2010), participants who had an error rate less than 2.5% (12 trials out of 480 trials) or greater than 30% (144 trials out of 480 trials) on either one of the sessions were excluded. The exclusion criteria of greater than 30% error trials were required to ensure that the participants were not performing at random (i.e., 50% is chance level). These

exclusion criteria are consistent with previous developmental studies (Davies et al., 2004; Gavin, et al., 2019).

This resulted in the loss of an additional 17 adults and 8 children for not making enough errors on either or both sessions, and 20 children due to making too many errors on either or both sessions. After imposing all of the exclusion criteria, data from 53 adults (M = 22.13 years, SD = 2.66) and 118 children (M = 10.19 years, SD = 1.47), were included for statistical analysis; see Table 1 for participants' age and sex distribution. According to Bujang & Baharum (2017, see Table 1a, p. 7), interclass correlation coefficient (ICC) statistical analyses to evaluate obtained R values of 0.35 or greater against hypothesized R_0 = 0 with power set to 80% and an alpha level at 0.05 requires a sample size of 46. Thus, the sample size after exclusions for each group is more than adequate for determining test-retest reliability in this current study.

Participants were compensated after each session with a choice of a cocoa mug, T-shirt, or cash, except for participants recruited from the Psychology department research subject pool who received course credits for participation. The study protocol was approved by the university institutional review board. Prior to study onset, all adult participants signed written consent forms, parents of child participants signed parental consent forms, and child participants signed assent forms.

Procedure

Participants were invited to the laboratory for two sessions scheduled 1 week apart though due to family schedules a few were schedule up to 3 weeks apart; 71% at 1 week, 25% at 2 weeks, 4% at 3 weeks. The short interval of 1-3 weeks was used to minimize the potential change in the brain development, a potential confounding variable in the reliability results, especially for children. Scheduling both visits on the same day of the week and at the same time of the day served as additional control for any potential confounding factors; e.g., biorhythms or daily activities differing from day to day during a week. Each visit included 1.5 hours of EEG tasks followed by 1 hour of behavioral testing with a 10-15 minute break between the EEG and behavioral testing. For the EEG portion, two trained research assistants prepped the participant for EEG recordings. After a 3-minute artifact training period, participants performed 3 separate ERP paradigms in a quiet recording room though only the results from the speeded visual flanker task are reported in this study. The behavioral testing included tasks of attention and executive function (these findings will be reported elsewhere) and were administered by a research assistant in another quiet testing area.

The ERP Paradigm

The speeded visual flanker task (Eriksen & Eriksen, 1974) was presented using E-prime software version 2.0 (Psychology Software Tools, Pittsburgh) in two blocks of 240 trials (480 trials total). In this task, participants were randomly presented four types of character arrays on the screen. Each character array consisted of combinations of the letters "H" or "S" organized as congruent arrays ("HHHHH" and "SSSSS", 80 trials each) and two incongruent arrays ("SSHSS" and "HHSHH", 160 trials each). Participants were instructed

to press either the left button on a 4 button keypad using their left index finger if the middle letter is an H and to press the right button using their right index finger if the middle letter is an S. Participants were told that the letters would be presented quickly, and they were instructed to perform as accurately as possible. The stimulus duration was 250 ms and the initial inter-stimulus interval (ISI) was set at 1400 ms. Following each set of 30 trials, the E-prime program was designed to evaluate the overall error rate and adjust the ISI by increasing or decreasing it by 100 ms if the error rate was greater than 30% or fewer than 10%, respectively. A minimal ISI was set at 800 ms to allow adequate time for brain processing of the stimulus and response to resolve prior to the onset of the stimulus on the subsequent trial. Behavioral measurements of error rate, response time (RT) on correct and incorrect trials were calculated for each of the two sessions.

Electrophysiological Recording

EEG data were collected from the scalp using either 33 channels or 64 channels from the same Active-Two BioSemi system (BioSemi, Inc., Amsterdam, the Netherlands) based on a modified 10-20 electrode placement system (American Electroencephalographic Society, 1994). Two electrodes, namely the common mode sense (CMS) and the driven right leg (DRL), were used to generate a common reference voltage (https://www.biosemi.com/faq/cms&drl.htm). Additional signals collected from the left and right earlobes were averaged and used for offline referencing. Two electrodes were placed at the supra- and infraorbital regions of the left eye to measure vertical eye movements. Two electrodes were placed at the left and right outer canthi to measure the horizontal eye movements. The sampling rate was 1024 Hz.

Electrophysiological Data Reduction

The EEG data were analyzed offline using Brain Vision Analyzer 2.0 software (www.brainproducts.com). The data were referenced to the averaged signals of bilateral earlobes and then filtered with a bandpass filter of 0.1–30 Hz with 24 dB/oct. The data were then segmented into response-locked and stimulus-locked segments.

For response-locked segments, the data on incorrect trials were segmented into 1400 ms time periods, which spanned from 600 ms before the incorrect response to 800 ms after the incorrect response. Segments with premature button responses (e.g. response times that were faster than 100 ms) were excluded from the analysis. Then, the segments were baseline-corrected based on the average voltage of −600 to 400 ms preceding the incorrect response (Davies, Segalowitz, & Gavin, 2004). Eye movement artifacts were removed via a regression approach based on the VEOG channel using customized MATLAB code (Segalowitz, 1996) then baseline-corrected again using the period of −600 to 400 ms preceding the incorrect response. Segments containing voltage greater than ± 100 μV in the midline (e.g. Fz, FCz, Cz, Pz) and VEOG channels were rejected. The segments were then averaged using traditional ERP data analysis and also processed with the Woody filter (defined below) then subsequently averaged after adjusting for latency jitter (see Table 2 for the average number of segments included in the averaged ERP by age and session). The windows for selecting the peaks for the ERN and Pe are reported below in the Adaptive Woody Filter section.

For stimulus-locked segments, the data were segmented into 1200 ms time periods, which spanned from 200 ms before stimulus onset to 1000 ms after stimulus onset. Then, the segments were baseline-corrected based on the average voltage of −200 to 0 ms of stimulus onset. Eye movement artifacts were removed via a regression approach based on the vertical EOG (VEOG) channel (Segalowitz, 1996) then baseline-corrected again using the period of −200 to 0 ms of stimulus onset. Segments containing voltage greater than ± 100 μV in the midline (e.g. Fz, FCz, Cz, Pz) and VEOG channels were rejected. The segments were then averaged to create averaged stimulus-locked ERPs for each participant. The stimulus-locked averaged ERPs obtained for each participant were scored using a customized peak-picking procedure programmed in MATLAB (Mathworks, Natick, MA). We used different time windows for measuring stimulus-locked ERPs in adults and children (Table 3), because two groups demonstrated different morphology of the ERP waveforms. The peaks were calculated based on the peak-to-peak measure. All of component were measured at the site FCz except for the P3 component was measured at Pz in addition to FCz. The topographic map that was used to determine the channel sites is presented in Figure 1.

Adaptive Woody Filter

After stimulus-locked ERP components were scored and stored in a database, the response-locked ERPs on the incorrect trials were processed using an Adaptive Woody filter programmed in MATLAB (Gavin et al., 2019; Luck, 2014; Woody, 1967). The individual files of response-locked segmented data were then passed to a custom-built MATLAB program that implements the Adaptive Woody Filter procedure to measure and adjust for the trial-to-trial variability in the latency of the ERN component. This template-matching procedure is described as follows. First, a template is obtained by averaging all the incorrect trials as in the traditional manner for generating a response-locked averaged ERP for the individual. Then, the maximal alignment of each segment is found through a three-step process. In the first step, the initial degree of alignment to the template is found by calculating a correlation coefficient for a data points in a window representing the 0 to 300 ms of the template. In the second step, the program searches for a better alignment *later* in the segment by shifting the segment data 120 times in increments of 1 data point (each data point = 0.98 ms) to the left. After each data point shift, the program again calculates a correlation coefficient between the shifted segment ERN waveform and the template ERN waveform for the 0 - 300 ms window to determine if the alignment produced a higher correlation value than previous shifts. Then starting over from the segment's original time relationship to the template, the third step searches for better alignment *earlier* in the segment by calculating the correlation again after each shifting of the segment data in increments of 1 data point to the right advancing up to 120 times or until the "N2" boundary is found, whichever was reached first. After all shifts were completed, the data point position that produced the highest correlation coefficient represents the amount of shift needed for each segment to maximally align with the template. When all segments are maximally aligned to the template window, the segments are then averaged to obtain a "latency-adjusted" averaged ERP waveform.

Using the time frame of 0 ms to 300 ms allows the Adaptive Woody Filter procedure to match segment morphology to the template morphology of the ERN waveform from the

preceding positivity of the P300 to about half of the Pe waveform, thereby enhancing the overall matching accuracy. However, without adding a safeguard, there exists a possibility that the shifting of the segment to the right to match earlier parts of the segment to the template window would allow the N2 component, along with portions of the P2 and P3, to be incorrectly chosen as the "ERN" in the segment. For example, if a participant had a segment with a reaction time of 300 ms, and the peak latency of his/her averaged N2 component was 200 ms, then it is likely that the program could potentially misidentify the N2 component as the ERN component when shifting the segment to the right is allowed to occur 120 times. To prevent the program from making a misidentification, the program determines a "N2" boundary *for each segment* to limit for sliding the window too early. This "N2" boundary is calculated by:

$$\text{Right shift limit} = (\text{response time for the segment in ms} - (\text{peak latency of the N2 derived from the stimulus-locked averaged ERP for the participant in ms} + 30\ \text{ms})) \ / \ (\text{A/D sampling interval in ms})$$

The value of 30 ms approximates the half cycle of the N2 component, and its addition to the peak-latency of the averaged N2 amplitude allows for a reasonable estimation of the boundary of the completed N2 cycle at the single segment level. Therefore, for the example above, based on the formula the limit of the number of points for shifting to the right for this individual segment is calculated as (300 – (200 + 30))/ (.97656) = 71.68 data points. Thus, the window for this segment is only allowed to be shifted to the right at a maximum of 71.68 data points instead of 120 data points.

After the data were processed through the Adaptive Woody filter, the latency-adjusted averaged ERP waveform were processed using the customized peak-picking program for scoring the ERN and Pe amplitudes at the FCz site. The window for measuring the ERN component was 10 ms prior to 180 ms after incorrect responses. The window for measuring the Pe component was 120 – 450 ms after incorrect responses. We used the same window for both children and adults. The peaks were calculated based on the peak-to-peak measure. The peak-to-peak measure was used because children often have an ERN deflection that is above baseline (with positive up) and the peak-to-peak measure preserves the ERN as a negative deflection (Gavin, et al., 2019). In a developmental study including children and adults of similar ages to those in this present study, Gavin and colleagues (2019) conducted split-half reliability of three measures of ERN amplitude; peak-to-peak, baseline to peak, and averaged ERN amplitude measures. They showed that the peak-to-peak measure resulted in the best internal consistency.

Statistical Analyses

The behavior measures, RT and error rate, were analyzed with the sole purpose of describing this study's samples to facilitate comparison with the previous studies. These descriptive analyses are not used to answer research questions and hypotheses. For overall behavioral outcomes, a three-way ANOVA was used to examine the effect of Group (Children and Adults), Session (Session 1 vs Session 2), and Trial Type (Correct vs Incorrect) on response times. A two-way ANOVA was used to investigate the effect of the Group (Children and Adults) and Session (Session 1 vs Session 2) on error rate. To answer our first research

question related to developmental differences, two three-way ANOVAs were used to examine the effect of Group (Children and Adults), Session (Session 1 vs Session 2), and Latency Jitter Correction (Before correction vs After correction) on the ERN amplitude and Pe amplitude, respectively. Paired-wise post hoc analyses were conducted using the pooled error term which according to Kirk (1968) takes into account the multiple comparisons. The corresponding p value of the observed studentized q value was calculated using the website (http://elvers.us/stats/tables/qprobability.html).

To answer our second and third questions Pearson correlations and two types of intraclass correlations (ICCs), ICC (3,1) consistency and ICC (3,1) absolute agreement, were used to assess the reliability of the ERN amplitude across two sessions for behavioral performances (e.g. response times and error rates), stimulus-locked ERPs (e.g. N1, P2, N2, P3), and response-locked ERPs (e.g. ERN, and Pe). The interpretation of the coefficients are as follows: <0.5 poor reliability; 0.5-0.75 moderate reliability; 0.75-0.90 good reliability; > 0.90 excellent reliability (Koo & Li, 2016; Portney & Watkins, 2009). Two online calculators were utilized to conduct the significance testing between correlation coefficients. Specifically, the correlation test for independent samples (http://www.quantpsy.org/corrtest/corrtest.htm) was used to compare the reliability of the ERPs obtaining across sessions between children and adults (Preacher, 2002); the correlation test for dependent correlations (http://www.quantpsy.org/corrtest/corrtest3.htm) was used to compare the reliability of the ERPs obtaining across sessions before and after latency jitter correlation for each group (Lee & Preacher, 2013). One-tail tests were used because our *a priori* hypotheses were directional; i.e. we predicted that the reliability of ERN and Pe amplitudes to be higher for adults compared to children, and that the reliability of ERN and Pe amplitudes to be higher after Woody filter adjustment compared to before adjustment.

Results

Descriptive Analyses

Behavioral results on the Flanker task.

Response times (RTs).: The descriptive results and the reliability of the RTs were presented in Table 4 and Figure 2. The three-way ANOVA demonstrated that the interaction between Group (children vs adults) x Session (session 1 vs session 2) x Trial Type (correct vs incorrect) on RTs was statistically significant, $F(1,169) = 9.22$, $p = .003$, $\eta_p^2 = 0.05$, as well as the main effect of the Group, $F(1,169) = 137.95$, $p < .001$, $\eta_p^2 = 0.45$. Post hoc analyses using the pooled error term demonstrated that the RTs in adults were always significantly faster than the RTs in children for any comparison of trial type/session combination ($p < .01$). Consistent with the literature, both adult and child groups demonstrated significantly faster RT for incorrect trials compared to correct trials in both sessions ($p < .01$; *one-tail*). Across sessions, adults generally show little change in RT for correct or incorrect trials. In contrast, and as a partial explanation for the 3-way interaction, children do show some slowing of RT of incorrect trials during session 2 ($M = 507$ ms) compared to session 1 ($M = 479$ ms), although not significant, while no considerable change in RT of correct trials during session 1 ($m = 624$ ms) and session 2 ($m = 620$ ms) was found.

Error rates.: The descriptive results and the reliability of the error rates are presented in Table 4 and Figure 3. The two-way ANOVA showed that the interaction between Group x Session reached statistical significance, $F(1, 169) = 16.33$, $p < .001$, $\eta_p^2 = 0.09$. Post hoc analyses showed that children made significantly more errors than adults on both sessions (ps < .001). For children, their error rate on session 1 was significantly greater than the error rate on session 2 ($p < .001$). however, for adults, no significant differences were found on the error rate across sessions (ns; $p > .05$).

Electrophysiological results.

Response-locked ERPs.: To determine if we could replicate developmental differences found by Davies et al. (2004), Gavin et al. (2019) and others in the ERN and Pe amplitude between adults and children using a speeded version of the flanker task, we conducted two three-way ANOVAs. The ANOVAs examined the effect of Group (Children vs Adults) x Session (Session 1 vs Session 2) x the Latency Jitter Correction (Before vs After) on the ERN and Pe amplitudes, respectively. The means and standard deviation of the ERN and Pe amplitudes and latencies before and after latency jitter correction for both sessions are reported in Tables 5 and 6. The ERPs are presented in Figure 4.

For the ERN amplitude, the main effect of the Group revealed that, overall the adults demonstrated significantly larger ERN amplitudes than children, $F(1,169) = 23.40$, $p < .001$, $\eta_p^2 = .12$ (Figure 5). The main effect for Latency Jitter Correction demonstrated that the ERN amplitude was larger after correction compared to before correction, $F(1,169) = 471.22$, $p < .0001$, $\eta_p^2 = .74$. The post hoc analyses confirmed that this was true for both groups and both sessions. The three-way interaction was also significant, $F(1,169) = 5.55$, $p = .02$, $\eta_p^2 = .03$. The pair-wise post hoc analyses comparing groups demonstrated that for session 1, the ERN amplitude was significantly larger for adults than children before latency jitter correction ($p < .05$), but the difference between adults and children decreased after the latency jitter correction (ns; $p > .05$). In contrast, for session 2 the mean ERN amplitude for adults was significantly larger than the mean amplitude for the children before and after the latency jitter correction ($p < .01$). For post hoc comparisons analyses comparing sessions, children showed only minor differences in the mean ERN amplitudes across sessions obtained before latency jitter correction, as well as after applying the correction (*ns*; $p > .05$). But, while adults showed an increased mean ERN amplitude from session 1 to session 2 on measures obtained before applying the latency jitter correction, it was not significant ($p > .05$). However, the average ERN amplitude change between sessions was significant for adults after the latency jitter correction ($p < .05$).

For the Pe amplitude, the main effect for Latency Jitter Correction demonstrated that the Pe amplitude was larger after correction compared to before correction, $F(1,169) = 921.98$, $p < .0001$, $\eta_p^2 = .85$ (Figure 6). The post hoc analyses confirmed that this was true for both groups and both sessions. The main effect for session was also significant, $F(1,169) = 23.11$, $p < .0005$, $\eta_p^2 = .12$, with the overall amplitude for session 2 being larger than session 1. In addition, both two-way interactions Group x Session, $F(1,169) = 6.81$, $p = .01$, $\eta_p^2 = .04$ and Group x Latency Jitter Correction $F(1,169) = 56.47$, $p < .001$, $\eta_p^2 = .25$ were significant. The pair-wise post hoc analyses for group differences showed that adults, as compared to

children, had significantly larger Pe amplitude on session 1 after latency jitter correction (p <.05), but the Pe amplitude was not significantly different between groups for session 1 before latency jitter correction (ns; p > .05). However, for session 2 the groups were not significantly different before or after latency correction (ns; p > .05).

Changes in latency jitter across sessions and groups.: The Adaptive Woody Filter used in this study obtains three measures of latency variability for each segment (Gavin, et al. 2019). Most relevant here is the shift value, a measure of the degree of time shifting, backward or forward, required to obtain the maximal correlation coefficient for the segment. To summarize the variability of the jitter latency across segments for an individual, a dependent measure, the $Shift_{SD}$, was computed as the standard deviation of the shift values across all segments for each participant.

To determine if the changes across sessions in ERN amplitude after the latency jitter correction were due to changes in latency jitter across sessions, the $Shift_{SD}$ values were evaluated using a 2 x 2 ANOVA with Group (Adults vs Children) was a between factor and Sessions (session 1 vs session 2) was a within factor. Across both groups, latency jitter decreased a small but significant amount from session 1 (M_{adults} = 49.36, $M_{children}$ = 63.71) to session 2 (M_{adults} = 46.54, $M_{children}$ = 62.74) with the main effect of session being statistically significant; $F(1,169) = 5.34$, $p = .022$, $\eta p^2 = .03$. Across both sessions, latency jitter was significantly greater for children compared to adults with the main effect of group being significant; $F(1,169) = 135.29$, $p < .0001$, $\eta p^2 = .45$. The interaction effect was not significant ($p = .26$). In addition, the within group variance for adults (SD_{s1} = 10.84, SD_{s2} = 11.47) was significantly greater than the within group variance for children (SD_{s1} = 7.42, SD_{s2} = 9.36); Levene's test of equality of error variance = 6.34, $p = .013$. Interestingly, the correlation of session 1 to session 2 was greater for adults ($r(53) = .52$, $p < .0001$) than for children ($r(118) = .39$, $p < .0001$).

Test-Retest Reliability Results

The ERN and Pe amplitudes before the latency jitter correction.—To answer our first research question, we conducted the Pearson correlation analyses and the ICC analyses to examine the test-retest reliability on the ERN and Pe amplitudes before the latency jitter correction. The results on the Pearson correlation showed that for adults, the reliability of the ERN and Pe amplitudes were .69 to .75, respectively. For children, the reliability of the ERN and Pe amplitudes were .55 and .62, respectively (Tables 5 and 6, Figure 7). In terms of the group differences on the reliability measures, the findings showed that before the latency jitter correction, the reliability of the ERN and Pe amplitudes in adults was not significantly higher than children (ERN: $z = 1.41$, $p = .08$ one tail; Pe: $z = 1.46$, $p = .07$ one tail).

The ERN and Pe amplitudes after the latency jitter correction.—To answer our second research question, we conducted the Pearson correlation analyses and the ICC analyses to examine the test-retest reliability on the ERN and Pe amplitudes after the latency jitter correction. The results showed that for adults, the reliability of the ERN and Pe amplitudes were .75 and .74, respectively. For children, the reliability of the ERN and Pe amplitudes were .57 and .65, respectively (Tables 5 and 6, Figure 7).

Contrary to what we hypothesized, the latency jitter correction did not significantly improve the reliability of the ERN and Pe amplitude in either adults and children (differences on the reliability of the ERN amplitude before and after Woody filter adjustment: adults: $z = -1.07$, $p = .14$, one tail, children: $z = -0.31$, $p = .38$, one tail; differences on the reliability of the Pe amplitude before and after Woody filter adjustment: adults: $z = 0.13$, $p = .44$, one tail, children: $z = -0.6$, $p = .27$, one tail). To examine the reliability in more detail, we broke down the child group by age and investigated the reliability of the ERN and Pe amplitude before and after latency jitter correction for each age group (Tables 5 and 6). The findings showed that for 8 year-old, and 10 year-old groups, the reliability of the ERN amplitude increased after the Woody filter adjustment, and these increments were statistical significant for 8 year-old ($z = -1.98$, $p = .02$, one tail) but not for 10 year-old (10 year-old: $z = -0.61$, $p = .27$, one tail). However, for groups of 9 year-old, 11 year-old, and 12 year-old, the ERN reliability decreased after the Woody filter, and the decrements were statistically significant for 12 year-old (12 year-old: $z = 1.91$, $p = 0.03$, one tail) but not for 9 and 11 year-old (9 year-old: $z = 0.12$, $p = .45$, one tail; 11 year-old: $z = 0.29$, $p = .38$, one tail).

In terms of the Pe, the reliability of the Pe amplitude increased after the Woody filter adjustment for 8 year-old, 9 year-old, 10 year-old, 12 year-old, yet none of these increments were statistically significant (8 year-old: $z = -0.85$, $p = .20$, one tail; 9 year-old: $z = -1.23$, $p = .11$, one tail; 10 year-old: $z = -0.90$, $p = .19$, one tail; 12 year-old: $z = -0.35$, $p = .36$, one tail). The reliability of the Pe amplitude decreased after the Woody filter adjustment for adults and 11 year olds, and the decrements were not statistically significant (11 year-old: $z = 0.64$, $p = .26$, one tail).

In terms of the group differences on the reliability measures, after the latency jitter correction, adults demonstrated significantly higher reliability of ERN amplitudes than children (z = 1.91, $p = .03$, one tail), but there were no significant differences before latency jitter ($z = 1.41$, $p = .08$, one tail). The reliability of the Pe amplitude for adults was not significantly higher than children either before or after latency jitter correction (before: Pe: $z = 1.46$, $p = .07$, one tail; after: $z = 1.07$, $p = .14$, one tail; see Figure 8).

Stimulus Time-locked test retest reliability.: We analyzed the reliability on the stimulus-locked ERPs (N1, P2, N2, P3) on correct and incorrect trials for contrastive purposes. The descriptive results and reliability indices of stimulus-locked ERPs of correct trials in session 1 and session 2 are reported in Tables 7 and 8. The ERPs are presented in Figure 6. Generally, the amplitude of stimulus-locked ERP components (N1, P2, N2, P3) were strongly correlated among sessions for correct trials for adults and children, (adults: $r_{min} = .78$, $r_{max} = .87$; children: $r_{min} = .71$, $r_{max} = .93$), but weaker for incorrect trials, (adults: $r_{min} = .32$, $r_{max} = .74$; children: $r_{min} = .37$, $r_{max} = .75$), especially for N1 amplitude.

Discussion

The present study examined the test-retest reliability of the ERN and Pe amplitudes before and after adjusting for the latency jitter in 53 neurotypical adults and 118 typically-developing children. For contrastive information, we also investigated the test-retest reliability of the mid-to-late ERP components elicited by the stimulus onset (i.e., stimulus

locked). We will discuss the results in terms of three aspects: the test-retest reliability of response-locked ERPs (ERN and Pe), the reliability of the stimulus-locked ERPs (N1, P2, N2, P3) on correct and incorrect trials, and the role of the latency jitter in the grand-averaged ERN and Pe amplitudes in adults and children.

The Test-retest Reliability of the ERN and Pe in Adults and Children

To our knowledge, this is the first study to examine the test-retest reliability of the ERN and Pe amplitudes over a short period (1- to 3-week) in typically-developing children. We found moderate reliability of the ERN and Pe amplitudes across sessions in typically-developing children aged 8-12 year-old (reliability of the ERN amplitude before latency jitter correction: $r = 0.55$, after latency jitter correction: $r = 0.57$; reliability of the Pe amplitude before latency jitter correction: $r = 0.62$, after latency jitter correction: $r = 0.65$). Previous studies on the test-retest reliability of the ERN amplitude have shown a general decreasing trend in the reliability of the ERN amplitude with increasing interval between sessions. However, the reliability of the ERN amplitude found in our study is slightly lower even with a shorter interval compared to the reliability of the ERN amplitude ($r = 0.63$) measured with a 2-year interval in children and adolescents aged 8 to13 years (Meyer et al., 2014).

There are several possible explanations for the discrepancy between the r value in the current study and the previous literature. First, our study required two visits with 1-3 weeks interval. Participants might have felt nervous (i.e., anxiety) at the first visit as the laboratory setting, equipment, procedures, and research assistants were novel to them. However, participants might have felt less nervous at the second visit. Whereas, in the Meyer et al. (2014) study that had a two-year interval between sessions, both sessions would more likely be a novel experience for the young children and the state effects may be more similar between the 2-year interval sessions compared to a 1-3 weeks interval in this present study. The differences in state across sessions may contribute to the unaccounted variance in the ERN and Pe measures which could lead to a lower test-retest reliability than previous studies. In addition to the state effects of anxiety, practice and learning strategies could also contribute to variable neural responses in the two sessions among children (Pauli et al., 1994; Romero, McFarland, Faust, Farrell, & Cacace, 2008; Taylor, Gavin, & Davies, 2016, Taylor, Gavin, Grimm, et al., 2019). Evidence for potential practice or learning effects in this current study is indicated by the lower error rates in session 2 compared to session 1. Second, this study utilized a different study design compared to Meyer et al. 2014. For example, the sample sizes (118 vs 44), paradigm (letter version vs arrowhead version), ISIs (1400ms +/− adjusted for error rate vs variable rate of 2300 – 2800 ms) and even task instructions in this study were different compared to Meyer et al., 2014, and these differences increase the difficulty of comparing the results.

Supporting these differences in reliabilities between studies described above, Clayson and Miller (2017), provide a comprehensive discussion of the importance of conducting and reporting psychometric data, both reliability and validity, in studies using event-related potential measures. Clayson and Miller emphasize that reliability is context dependent, and reliability reported in one study may not directly apply to another study, especially if the population, paradigm, data processing stream, and other parameters of the studies differ.

Thus, the differences in reliability reported in various studies reviewed above is not surprising. Due to reliability being context dependent, Clayson and Miller recommend that reliability should be reported for all studies. These authors also discussed the role of reducing noise and measurement error for producing more reliable ERP measures. Many aspects of a study can contribute to measurement error and it is important for a study to control measurement error, which can be especially challenging when collecting ERP data in children and clinical populations. Gavin and Davies (2008) suggest both procedural considerations as well as a statistical model for possible sources of variance and ways of minimizing measurement error. In support of their theoretical/statistical model of controlling for multiple sources of variance, Taylor, et al. (2019) using structural equation modeling demonstrated that individual emotional or physical state accounted for a significant amount of variance in ERP measurements. In addition, individual differences in states varied systematically across two sessions, possibly representing changes in anxiety, learning or shifts in cognitive strategies from one session to the next.

Consistent with previous literature, our findings with neurotypical adults demonstrated moderate to strong test-retest reliability of the ERN and Pe amplitudes (r_{min} = .69 to r_{max} =.75; Segalowitz et al., 2010; Olvet & Hajcak, 2009a; Cassidy et al., 2012; Weinberg et al., 2011). Taken together, the findings suggest that the ERN and Pe amplitudes are reliable measures across time for both adults and children. However, for clinical diagnostic purposes, reliability measures exceeding .80 are more desirable (Nunnally, 1978, p. 245). Given that studies are not yet demonstrating this level, there are still other sources of uncontrolled variance (e.g. state effect) that need to be considered if researchers are to establish ERN and Pe amplitudes as biomarkers for neurological disorders. For example, studies have shown that the ERN component is influenced by state effects such as fatigue (Lorist, Boksem, & Ridderinkhof, 2005), or sleep deprivation (Scheffers et al., 1999), such that people with sleep deprivation have a smaller ERN and Pe amplitudes (Tsai, Young, Hsieh, & Lee, 2005). The majority of the adult participants in our sample are graduate or undergraduate students who were evaluated during the school year. Although we scheduled their visits on the same time of the day during the same day of the week, we did not take the level of the sleep deprivation into the consideration (e.g. if the session was scheduled on the same day when the participant had mid-term or final tests, he/she might have stayed up late the previous night). Moreover, the potential session effect (e.g. practice effect) should also need to be taken into consideration when measuring the test-retest reliability in adults and children. Future studies should be conducted to understand the extent the state effect influences the test-retest reliability of the ERN and Pe amplitudes.

The number of trials used to produce averaged ERN and Pe amplitudes may also influence the test-retest reliability (Clayson & Miller, 2017; Larson et al., 2010; Baldwin et al., 2015; Olvet & Hajcak, 2009b; Pontifex et al., 2010). However, in this current study there was little to no change in the reliability when partialling out the number of trials in regression analyses. For children, the reliability increased by .001 for ERN amplitudes obtained before latency jitter adjustment and there was no change for ERN obtained after latency jitter adjustment. For adults, the reliability increased by .024 for ERN amplitudes before latency jitter adjustment and increased by .014 for ERN after latency jitter adjustment. For children after partialling out the number of trials, the Pe amplitudes reliability increased by .007

before latency jitter adjustment and no change in reliability after jitter adjustment. For adults, the reliability of Pe amplitudes increased by .001 before latency jitter adjustment and increased by .001 after latency jitter adjustment. Therefore, in this current study the number of trials had little to no effect on reliability.

This finding that the number of trials (segments) had little to no effect on the test-retest reliability in this current study is not surprising because of the two methodological approaches employed to control for this. The first approach used was the criteria for inclusion and exclusion of participants based on number of error trials. The requirement of 12 error trials (2.5% or greater error rate) for the participant's data to be included in the data analyses was chosen because an averaged ERP with 12 trials will have less variability in the ERN and Pe measures due to a better signal-to-noise ratio (Clayson & Miller, 2017). Thus, an ERN measure taken from an averaged ERP that includes at least 12 trials will be a better estimate of that individual's ERN than if the measurement was obtained from an averaged ERPs with only 6 – 11 trials. There are only two other studies examining ERN in children while using the Woody Filter and one required at least 2% error rate (Tabachnick, et al., 2018) and the other required 2.5% error rate (Gavin, et al., 2019). The exclusion criteria of greater than 30% error trials were required to ensure that the participants were not performing at random (i.e., 50% is chance level).

The second methodological approach used in this current study was the adjustment of the speed of the stimulus presentation (i.e., changing the inter-stimulus interval) in the flanker paradigm based on the cumulative error-rate that help control for the number of error trials. In this approach the error rate was evaluated after every 30 trials and the ITI was adjusted up or down if needed to keep error rate within 10% to 30%. The only other ERN developmental study that attempted to maintain a certain error rate level (Hannah, et al., 2012) provided feedback at the end of a block of 32 trials to encourage more accuracy or faster responses to keep the error rates close to 10%. The age of participants in the Hannah et al. study was 10 – 19 years and the averaged error rates were between 10% – 12%. In the current study with slightly younger children starting at age 8, the averaged error rates were very similar to the Hannah et al. study; i.e., 14.79% (approximately 70 trials) for session one, and 10.73% for session two (approximately 51 trials). Future studies examining the effect of number of trials on obtaining stable test-retest reliability are warranted, particularly for those involving pediatric populations.

Reliability of the Stimulus-Locked ERPs

We analyzed the reliability on the stimulus-locked ERPs (N1, P2, N2, P3) on correct and incorrect trials for contrastive purposes. As expected, the test-retest reliability is good on the stimulus-locked ERPs for correct trials (adults: r_{min} = .78, r_{max} = .87; children: r_{min} = .71, r_{max} = .93), but relatively poor for the incorrect trials (adults: r_{min} = .32, r_{max} = .74; children: r_{min} = .37, r_{max} = .75). We will discuss these findings in terms of two aspects. First, the good test-retest reliability on the stimulus-locked ERPs especially for N1 on correct trials for both adults (r_{N1} = 0.78) and children (r_{N1} = 0.83) speaks to our task validity such that participants generally attended to the stimuli in the task across sessions. Specifically, the N1 component has been associated with selective attention and the early stimulus discrimination process

and is larger for attended stimuli compared to ignored stimuli (Lackner, Santesso, Dywan, Wade, & Segalowitz, 2013; Luck & Girelli, 1998; Polich, 1993). Had participants not attended to the task stimuli consistently across the sessions, we would not have obtained the good reliability on the N1 amplitude. The good test-retest reliability on the stimulus-locked ERPs implies that the relatively less strong test-retest reliability on the response-locked ERPs (i.e. ERN and Pe) is not due to a lack of attention to the task in general. It could be that the response-locked ERPs involved with more endogenous cognitive processes, such as error detection, could have been influenced by motor responses such as button presses (Ullsperger & Von Cramon, 2001). Second, the reliability on the stimulus-locked ERPs on the incorrect trials is lower compared to that of the correct trials. These results suggest a lapse in the attention to the stimulus—which, in turn, led to a failure to inhibit the prepotent motor response (i.e., button press) and an insufficient processing of the stimuli, and subsequently, resulting in an incorrect button presses (van Veen and Carter et al. 2006).

Latency Jitter as a Trait-like Measure

Contrary to our hypothesis, our findings demonstrated that the test-retest reliability of the ERN and Pe amplitudes were not significantly improved after adjusting for the latency jitter in adults and all ages of children. One possible explanation could be that the latency jitter may be a trait-like variable and consistently occurred across the two sessions. As a result, removing the latency jitter across two sessions did not improve the test-retest reliability of the ERN and Pe amplitudes. Several studies support the notion that the latency jitter is a trait-like measure whereby individuals with certain traits have greater amount of latency jitter. For instance, the latency jitter has been shown to be greater in people with schizophrenia compared to their neurotypical peers (Young et al., 2001) and in older adults compared to young adults (McDowell, Kerick, Santa Maria, & Hatfield, 2003). Moreover, in a study that employed a three-stimuli visual oddball task with young to elderly adults aged 20 to 89 years, latency jitter of the P3a component was shown to be correlated with age (r = 0.28, p < .002; Fjell, Rosquist, & Walhovd, 2009). Additionally, latency jitter may be an indicator of processing efficiency (i.e. the efficiency of the neural transduction) and has been associated with other trait measures such as working memory (Shucard, Covey, & Shucard, 2016) and shifting/inhibition (Fjell et al., 2009). It is unlikely that the neural systems or cognitive functions underwent drastic changes during our experimental period (i.e., 1-3 weeks), and, as a result, it is understandable that adjusting for the latency jitter did not improve the reliability of the ERN and Pe amplitude. Furthermore, previous studies have suggested that the intra-individual variability in response time on behavioral tasks as a trait-like measure (Hultsch, Hunter, MacDonald, & Strauss, 2005). While the latency jitter correction did not improve the reliability of the ERN and Pe amplitudes in children and adults as a group, it is worth noting that when we examined the reliability of these ERP components for each age group in children, the reliability of the ERN amplitude significantly improved for age group 8 and significantly decreased for age group 12 after adjusting the latency jitter. Since the groups of 8 year-olds and 12 year-olds are the lower and upper bound of participant's age range in this study, the findings suggest that the latency jitter could reflect different neural processing characteristics for younger or older children. Notably, a recent study demonstrated that latency jitter changed as a linear function of age (Gavin, et al., 2019), adding to the evidence that the latency jitter trait may reflect neural

development. Future studies are needed to investigate the differential effect of correcting for latency jitter across development including younger and older age groups.

The age and latency jitter variables are correlated, and there is collinearity between these two variables; thus is it difficult to determine unique contributions of these two variables in ERN and Pe amplitude across development. However, both variables could reflect brain maturation. As mentioned above, other research groups have suggested that latency jitter may be related to working memory (Shucard, et al., 2016) and shifting/inhibition (Fjell et al., 2009). We propose that the latency jitter $Shift_{SD}$ may represent cognitive control or attentional control. Individuals that have more cognitive or attentional control will more likely have consistent timing of brain processing of stimuli during the ERN task from trial to trial and consequently, have smaller latency shifts and smaller $Shift_{SD}$. In adults who have more mature brains than children, we would expect more consistency in the timing of processing reflecting better attention and cognitive control within the task at hand. Indeed, the mean $Shift_{SD}$ for adults ($M_{session1}$ = 49.36, $M_{session2}$ = 46.54) are smaller than that of the children ($M_{session1}$ = 63.71, $M_{session2}$ = 62.74). Furthermore, when we correlated the latency jitter across sessions, we found that the adults have greater consistency of $Shift_{SD}$ from session 1 to session 2 (r = .52) than children (r = .31). Consequently, the more consistent timing of processing across sessions in adults as shown with the $Shift_{SD}$ measure, suggests a more established trait behavior over time compared to the children. Given that the correlation for the child group across sessions was also significant but smaller, this suggests that this trait may not be as established. Alternately, state factors such as anxiety, learning and developing cognitive strategies may be confounding the results making the interpretation more challenging in children than adults.

Conclusion

We found moderate to strong test-retest reliability of the ERN and Pe amplitudes in neurotypical adults and moderate test-retest reliability of the ERN and Pe amplitudes in typically-developing children aged 8 to 3 years with a 1-3 weeks interval between the sessions. However, contrary to our hypothesis, the test-retest reliability did not improve after the latency jitter correction, suggesting that latency jitter may not markedly contribute to variance across sessions. Additionally, the stimulus-locked ERPs on correct trials demonstrated strong reliability in children and adults, ruling out the possibilities that variant attention levels on the task stimuli across sessions caused a lower reliability of the ERN and Pe amplitudes compared to previous studies. Future studies could explore other factors such as controlling for state effects that may enhance the psychometric properties of the ERN and Pe amplitudes in children and adults.

Acknowledgements:

This study was funded in part by NIH/NICHD (R03HD046512) to PLD and WJG, the Sigma XI Research Foundation to MHL, the Colorado State University College of Health and Human Sciences, and the Colorado State University Departments of Occupational Therapy, and Human Development and Family Studies. These data were analyzed as a portion of a doctoral dissertation effort.

References

American Electroencephalographic Society. (1994). Guideline thirteen: Guidelines for standard electrode position nomenclature. Journal of Clinical Neurophysiology, 11(1), 111–3. [PubMed: 8195414]

Baldwin SA, Larson MJ, & Clayson PE (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. Psychophysiology, 52, 790–800. doi:10.1111/psyp.12401 [PubMed: 25581577]

Bates AT, Liddle PF, Kiehl KA, & Ngan ETC (2004). State dependent changes in error monitoring in schizophrenia. Journal of Psychiatric Research, 38(3), 347–356. 10.1016/j.jpsychires.2003.11.002 [PubMed: 15003441]

Carrasco M, Harbin SM, Nienhuis JK, Fitzgerald KD, Gehring WJ, & Hanna GL (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and unaffected siblings. Depress Anxiety, 30(1), 39–46. 10.1002/da.22035 [PubMed: 23225541]

Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, & Cohen JD (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. Science, 280(5364), 747–749. doi:10.1126/science.280.5364.747 [PubMed: 9563953]

Cassidy SM, Robertson IH, & O'Connell RG (2012). Retest reliability of event-related potentials: evidence from a variety of paradigms. Psychophysiology, 49(5), 659–664. 10.1111/j.1469-8986.2011.01349.x [PubMed: 22335452]

Clayson PE, & Miller GA (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. International Journal of Psycholophysigology, 111, 57–67. 10.1016/j.ijpsycho.2016.10.012

Coles MGH, Scheffers MK, & Holroyd CB (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. Biological Psychology, 56(3), 173–189. 10.1016/s0301-0511(01)00076-x [PubMed: 11399349]

Davies PL, Segalowitz SJ, Dywan J, & Pailing PE (2001). Error-negativity and positivity as they relate to other ERP indices of attentional control and stimulus processing. Biological Psychology, 56(3), 191–206. doi:10.1016/S0301-0511(01)00080-1 [PubMed: 11399350]

Davies PL, Segalowitz SJ, & Gavin WJ (2004). Development of error-monitoring event-related potentials in adolescents. Annals of the New York Academy of Sciences, 1021, 324–328. 10.1196/annals.1308.039 [PubMed: 15251904]

DuPuis D, Ram N, Willner CJ, Karalunas S, Segalowitz SJ, & Gatzke-Kopp LM (2014). Implications of ongoing neural development for the measurement of the error-related negativity in childhood. Developmental Science. 10.1111/desc.12229

Falkenstein M, Hohnsbein J, Hoormann J, & Blanke L (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. Electroencephalography and Clinical Neurophysiology, 78(6), 447–455. doi:10.1016/0013-4694(91)90062-9 [PubMed: 1712280]

Falkenstein M, Hoormann J, Christ S, & Hohnsbein J (2000). ERP components on reaction errors and their functional significance: a tutorial. Biological Psychology, 51(2-3), 87–107. [PubMed: 10686361]

Fjell AM, Rosquist H, & Walhovd KB (2009). Instability in the latency of P3a/P3b brain potentials and cognitive function in aging. Neurobiology of Aging, 30(12), 2065–2079. 10.1016/j.neurobiolaging.2008.01.015 [PubMed: 18339453]

Foti D, Kotov R, & Hajcak G (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. Journal of Abnormal Psychology, 122(2), 520–531. 10.1037/a0032618 [PubMed: 23713506]

Gavin WJ, & Davies PL (2008). Obtaining reliable psychophysiological data with child participants: Methodological considerations. In Schmidt LA & Segalowitz SJ (Eds.), Developmental Psychophysiology: Theory, Systems, and Methods (pp. 424–447). New Tork, NY: Cambridge University Press.

Gavin WJ, Lin M-H, & Davies PL (2019) Developmental Trends of Performance Monitoring Measures in 7-to 25-year-olds: Unravelling the Complex Nature of Brain and Behavioral Measures. Psychophysiology, 56:e13365. 10.1111/psyp.13365. [PubMed: 30942480]

Hanna GL, Carrasco M, Habvin SM, Neinhuis JK, LaRosa CE, Chen P, Fitzgerald KD, Gehring WJ (2012). Error-related negativity and tic history in pediatric obsessive-compulsive disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 51, 902–910. doi: 10.1016/j.jaac.2012.06.019 [PubMed: 22917203]

Herrmann MJ, Römmler J, Ehlis AC, Heidrich A, & Fallgatter AJ (2004). Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe). Cognitive Brain Research, 20(2), 294–299. 10.1016/j.cogbrainres.2004.02.013 [PubMed: 15183400]

Holroyd CB, Dien J, & Coles MGH (1998). Error-related scalp potentials elicited by hand and foot movements: evidence for an output-independent error-processing system in humans. Neuroscience Letters, 242(2), 65–68. doi:10.1016/S0304-3940(98)00035-4 [PubMed: 9533395]

Hultsch DF, Hunter MA, MacDonald SWS, & Strauss E (2005). Inconsistency in Response Time as an Indicator of Cognitive Aging. In Duncan J, Phillips L, and McLeod P (Eds.), Measuring the Mind: Speed, Control, and Age (pp.32–57). Oxford, UK: Oxford University Press.

Kim MS, Seung SK, Kyung SS, So YY, Young YK, & Jun SK (2006). Neuropsychological correlates of error negativity and positivity in schizophrenia patients. Psychiatry and Clinical Neurosciences, 60(3), 303–311. 10.1111/j.1440-1819.2006.01506.x [PubMed: 16732746]

Kirk RE (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Koo TK, & Li MY (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15, 155–163. doi: 10.1016/j.jcm.2016.02.012 [PubMed: 27330520]

Lackner CL, Santesso DL, Dywan J, Wade TJ, & Segalowitz SJ (2013). Electrocortical indices of selective attention predict adolescent executive functioning. Biological Psychology, 93(2), 325–333. doi:10.1016/j.biopsycho.2013.03.001 [PubMed: 23528784]

Ladouceur CD, Dahl RE, Birmaher B, Axelson DA, & Ryan ND (2006). Increased error-related negativity (ERN) in childhood anxiety disorders: ERP and source localization. Journal of Child Psychology and Psychiatry and Allied Disciplines, 47(10), 1073–1082. 10.1111/j.1469-7610.2006.01654.x

Larson MJ, Baldwin SA, Good DA, & Fair JE (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials. Psychophysiology, 47, 1167–1171. doi:10.1111/j.1469-8986.2010.01022.x [PubMed: 20477982]

Larson MJ, Clayson PE, & Clawson A (2014). Making sense of all the conflict: a theoretical review and critique of conflict-related ERPs. International Journal of Psychophysiology, 93(3), 283–297. doi:10.1016/j.ijpsycho.2014.06.007 [PubMed: 24950132]

Larson MJ, Kaufman DAS, Kellison IL, Schmalfuss IM, & Perlstein WM (2009). Double Jeopardy! The Additive Consequences of Negative Affect on Performance-Monitoring Decrements Following Traumatic Brain Injury. Neuropsychology, 23(4), 433–444. 10.1037/a0015723 [PubMed: 19586208]

Lee IA, & Preacher KJ (2013, 10). Calculation for the test of the difference between two dependent correlations with no variable in common [Computer software]. Available from http://quantpsy.org.

Lorist MM, Boksem MAS, & Ridderinkhof KR (2005). Impaired cognitive control and reduced cingulate activity during mental fatigue. Cognitive Brain Research, 24(2), 199–205. 10.1016/j.cogbrainres.2005.01.018 [PubMed: 15993758]

Luck SJ (2014). An introduction to the event-related potential technique. (2nd ed.). Cambridge, MA: MIT Press.

Luck SJ, & Girelli M (1998). Electrophysiological approaches to the study of selective attention in the human brain. In Parasuraman R (Ed.), The Attentive Brain (pp. 71–94). Cambridge, MA, US: The MIT Press.

Lukie CN, Montazer-Hojat S, & Holroyd CB (2014). Developmental changes in the reward positivity: an electrophysiological trajectory of reward processing. Developmental Cognitive Neuroscience, 9, 191–199. 10.1016/j.dcn.2014.04.003 [PubMed: 24879113]

Mathalon DH, Whitfield SL, & Ford JM (2003). Anatomy of an error: ERP and fMRI. Biological Psychology, 64(1-2), 119–141. [PubMed: 14602358]

McDowell K, Kerick SE, Santa Maria DL, & Hatfield BD (2003). Aging, physical activity, and cognitive processing: An examination of P300. Neurobiology of Aging, 24(4), 597–606. 10.1016/S0197-4580(02)00131-8 [PubMed: 12714117]

Meyer A (2017). A biomarker of anxiety in children and adolescents: A review focusing on the error-related negativity (ERN) and anxiety across development. Developmental Cognitive Neuroscience, 27, 58–68. 10.1016/j.dcn.2017.08.001 [PubMed: 28818707]

Meyer A, Bress J, & Proudfit GH (2014). Psychometric properties of the error-related negativity in children and adolescents. Psychophysiology, 51(7), 602–610. [PubMed: 24646380]

Morris SE, Heerey EA, Gold JM, & Holroyd CB (2008). Learning-related changes in brain activity following errors and performance feedback in schizophrenia. Schizophrenia Research, 99(1–3), 274–285. 10.1016/j.schres.2007.08.027 [PubMed: 17889510]

Morris SE, Yee CM, & Nuechterlein KH (2006). Electrophysiological analysis of error monitoring in schizophrenia. Journal of Abnormal Psychology, 115(2), 239–250. 10.1037/0021-843X.115.2.239 [PubMed: 16737389]

Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GP, & Kok A (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology, 38(5), 752–760. [PubMed: 11577898]

Nunnally JC (1978). Psychometric Theory (2nd ed.). NewYork, NY: McGraw-Hill.

Olvet DM, & Hajcak G (2008). The error-related negativity (ERN) and psychopathology: toward an endophenotype. Clinical Psychology Review, 28(8), 1343–1354. doi:10.1016/j.cpr.2008.07.003 [PubMed: 18694617]

Olvet DM, & Hajcak G (2009a). Reliability of error-related brain activity. Brain Research, 1284, 89–99. 10.1016/j.brainres.2009.05.079 [PubMed: 19501071]

Olvet DM, & Hajcak G (2009b). The stability of error-related brain activity with increasing trials. Psychophysiology, 46, 957–961. doi:10.1111/j.1469-8986.2009.00848.x [PubMed: 19558398]

Overbeek TJM, Nieuwenhuis S, & Ridderinkhof KR (2005). Dissociable Components of Error Processing. Journal of Psychophysiology, 19(4), 319–329. 10.1027/0269-8803.19.4.319

Pauli P, Lutzenberger W, Rau H, Birbaumer N, Rickard TC, Yaroush RA, & Bourne LE (1994). Brain potentials during mental arithmetic: effects of extensive practice and problem difficulty. Cognitive Brain Research, 2(1), 21–29. 10.1016/0926-6410(94)90017-5 [PubMed: 7812175]

Polich J (1993). Cognitive Brain Potentials. Current Directions in Psychological Science, 2(6), 175–179. 10.1111/1467-8721.ep10769728

Pontifex MB, Scudder MR, Brown ML, O'Leary KC, Wu CT, Themanson JR, & Hillman CH (2010). On the number of trials necessary for stabilization of error-related brain activity across the life span. Psychophysiology, 47, 767–773. doi:10.1111/j.1469-8986.2010.00974.x [PubMed: 20230502]

Portney LG, Watkins MP. (2009) Foundations of clinical research: applications to practice (3rd ed). Upper Saddle River, New Jersey: Pearson Education, Inc.

Preacher KJ (2002, 5). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Available from http://quantpsy.org.

Rabella M, Grasa E, Corripio I, Romero S, Mañanas MÀ, Antonijoan RM, … Riba J (2016). Neurophysiological evidence of impaired self-monitoring in schizotypal personality disorder and its reversal by dopaminergic antagonism. NeuroImage: Clinical, 11, 770–779. 10.1016/j.nicl.2016.05.019 [PubMed: 27330977]

Ridderinkhof KR, Ramautar JR, & Wijnen JG (2009). To P(E) or not to P(E): a P3-like ERP component reflecting the processing of response errors. Psychophysiology, 46(3), 531–538. doi:10.1111/j.1469-8986.2009.00790.x [PubMed: 19226310]

Riesel A, Weinberg A, Endrass T, Meyer A, & Hajcak G (2013). The ERN is the ERN is the ERN? Convergent validity of error-related brain activity across different tasks. Biological Psychology, 93(3), 377–385. 10.1016/j.biopsycho.2013.04.007 [PubMed: 23607999]

Romero SG, McFarland DJ, Faust R, Farrell L, & Cacace AT (2008). Electrophysiological markers of skill-related neuroplasticity. Biological Psychology, 78(3), 221–230. 10.1016/j.biopsycho.2008.03.014 [PubMed: 18455861]

Ruchsow M, Herrnberger B, Beschoner P, Grön G, Spitzer M, & Kiefer M (2006). Error processing in major depressive disorder: Evidence from event-related potentials. Journal of Psychiatric Research, 40(1), 37–46. 10.1016/j.jpsychires.2005.02.002 [PubMed: 15882872]

Segalowitz SJ (1996). EYEREG.EXE program for epoch-based eye-channel correction of ERPs. St. Catherines, Canada: Brock University.

Segalowitz SJ, & Dywan J (2009). Individual differences and developmental change in the ERN response: implications for models of ACC function. Psychological Research, 73(6), 857–870. 10.1007/s00426-008-0193-z [PubMed: 19023593]

Segalowitz SJ, Santesso DL, Murphy TI, Homan D, Chantziantoniou DK, & Khan S (2010). Retest reliability of medial frontal negativities during performance monitoring. Psychophysiology, 47(2), 260–270. 10.1111/j.1469-8986.2009.00942.x. [PubMed: 20030755]

Scheffers MK, & Coles MG (2000). Performance monitoring in a confusing world: Error related brain activity, judgments of response accuracy, and types of errors. Journal of Experimental Psychology: Human Perception and Performance, 26, 141–151. [PubMed: 10696610]

Scheffers MK, Humphrey DG, Stanny RR, Kramer AF, & Coles MGH (1999). Error-related processing during a period of extended wakefulness. Psychophysiology, 36(2), 149–157. 10.1017/S0048577299980307 [PubMed: 10194961]

Shucard DW, Covey TJ, & Shucard JL (2016). Single trial variability of event-related brain potentials as an index of neural efficiency during working memory. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9743, 273–283. 10.1007/978-3-319-39955-3_26

Tabachnick AR, Valadez EA, Palmwood EN, Zajac L, Simons RF, & Dozier M (2018). Depressive symptoms and error-related brain activity in CPS-referred children. Psychophysiology, 55(11): e13211. doi:10.1111/psyp.13211. [PubMed: 30094846]

Taylor BK, Gavin WJ, & Davies PL (2016). The Test-Retest Reliability of the Contingent Negative Variation (CNV) in Children and Adults. Developmental Neuropsychology, 41, 162–175. doi: 10.1080/87565641.2016.1170835 [PubMed: 27145115]

Taylor BK, Gavin WJ, Grimm K, Prince MA, Lin M-H, & Davies PL (2019). A Unified Model of Event-Related Potentials as Phases of Stimulus-to-Response Processing. Neuropsychologia, 132. 10.1016/j.neuropsychologia.2019.107128

Tsai LL, Young HY, Hsieh S, & Lee CS (2005). Impairment of error monitoring following sleep deprivation. Sleep, 28(6), 707–713. 10.1093/sleep/28.6.707 [PubMed: 16477957]

Ullsperger M, & Von Cramon DY (2001). Subprocesses of performance monitoring: A dissociation of error processing and response competition revealed by event-related fMRI and ERPs. NeuroImage, 14(6), 1387–1401. 10.1006/nimg.2001.0935 [PubMed: 11707094]

Unsal A, & Segalowitz SJ (1995). Sources of P300 attenuation after head injury: single-trial amplitude, latency jitter, and EEG power. Psychophysiology, 32(3), 249–256. [PubMed: 7784533]

Van Boxtel GM (1998). Computational and statistical methods for analyzing event-related potential data. Behavior Research Methods, Instruments, & Computers, 30(1), 87–102. 10.3758/BF03209419

Van De Voorde S, Roeyers H, & Wiersema JR (2010). Error monitoring in children with ADHD or reading disorder: An event-related potential study. Biological Psychology, 84(2), 176–185. 10.1016/j.biopsycho.2010.01.011 [PubMed: 20097256]

van Veen V, & Carter C (2002). The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiology & Behavior, 77(4-5), 477–482. doi:10.1016/S0031-9384(02)00930-7 [PubMed: 12526986]

van Veen V, & Carter CS (2006). Error detection, correction, and prevention in the brain: a brief review of data and theories. Clinical EEG and Neuroscience, 37(4), 330–335. [PubMed: 17073172]

Weinberg A, & Hajcak G (2011). Longer term test-retest reliability of error-related brain activity. Psychophysiology, 48(10), 1420–1425. 10.1111/j.1469-8986.2011.01206.x [PubMed: 21496055]

Weinberg A, Dieterich R, & Riesel A (2015). Error-related brain activity in the age of RDoC: A review of the literature. International Journal of Psychophysiology, 98, 276–299. doi:10.1016/j.ijpsycho.2015.02.029 [PubMed: 25746725]

Wessel J (2012). Error awareness and the error-related negativity: Evaluating the first decade of evidence. Frontiers in Human Neuroscience, 6, 88. [PubMed: 22529791]

Woody CD (1967). Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Medical & Biological Engineering, 5, 539–553.

Yeung N, Botvinick MM, & Cohen JD (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. Psychological Review, 111(4), 931–959. 10.1037/0033-295x.111.4.939 [PubMed: 15482068]

Young KA, Smith M, Rawls T, Elliott DB, Russell IS, & Hicks PB (2001). N100 evoked potential latency variation and startle in schizophrenia. NeuroReport, 12(4), 767–773. 10.1097/00001756-200103260-00031 [PubMed: 11277581]

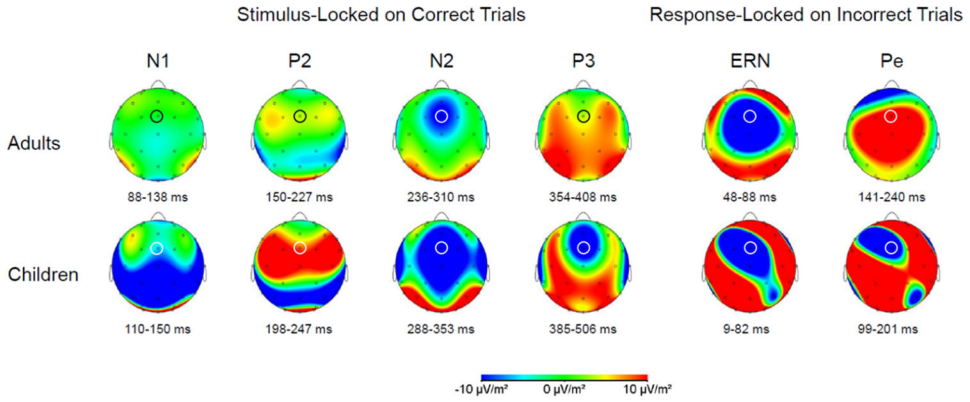

Figure 1.

The topographic distribution on the stimulus-locked correct trials (N1, P2, N2, and P3 components) and response-locked incorrect trials (ERN and Pe components) for adults and children. The channel FCz is marked with a while/black circle. Note: the time windows used to determine the topographic distribution for each component were calculated based on the averaged mean latency across sessions +/− averaged mean standard deviation across sessions for each age group

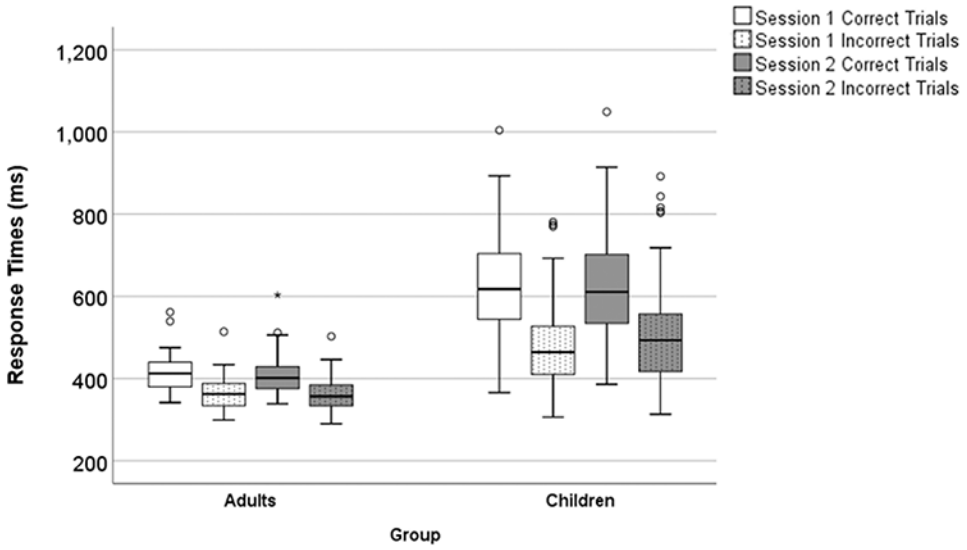

Figure 2.

The boxplot of response times on correct and incorrect trials for session 1 and session 2 in children and adults

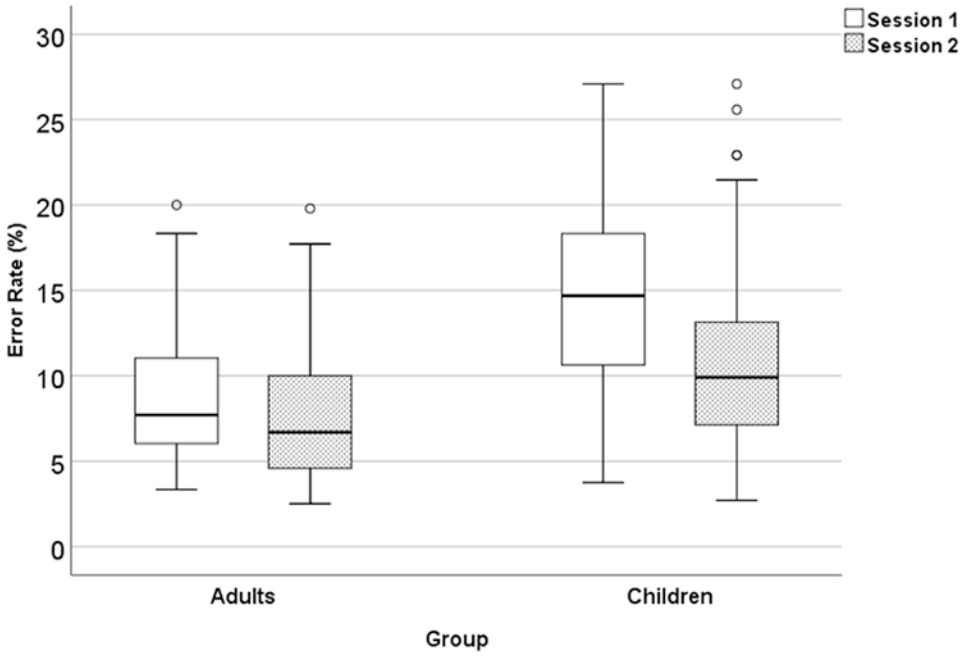

Figure 3.
The boxplot of error rates for session 1 and session 2 in children and adults

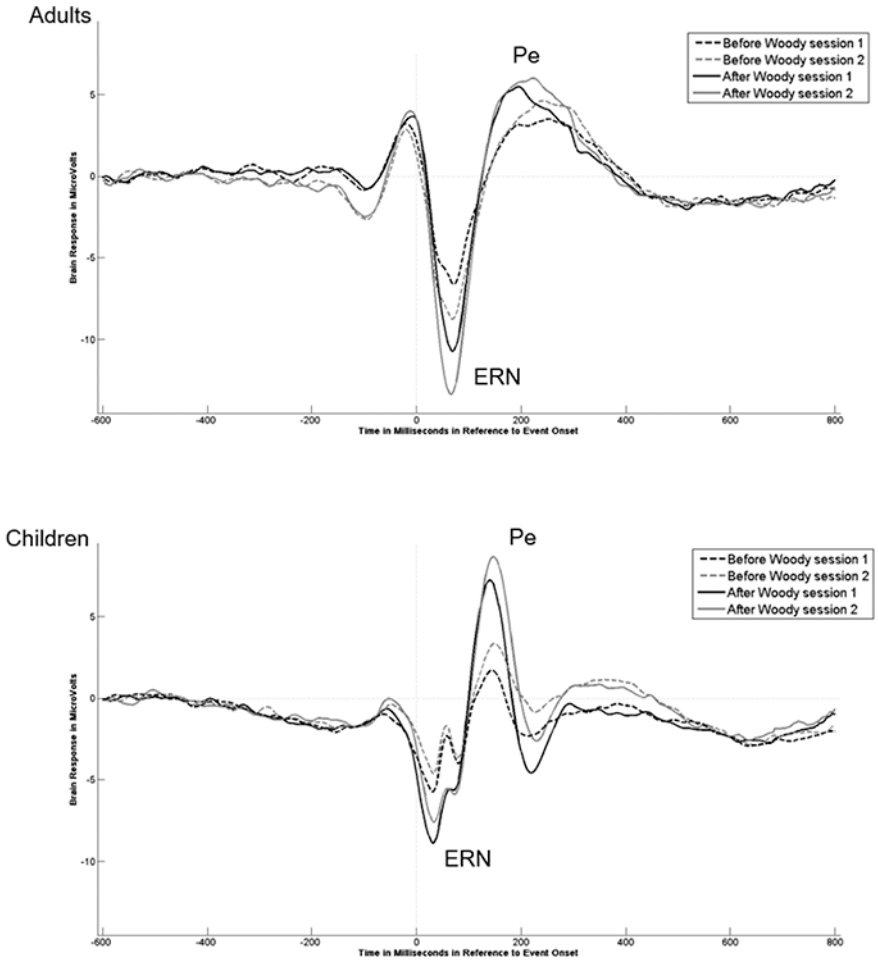

Figure 4.

The ERN amplitude for session 1 and session 2 from the channel of FCz before and after latency jitter correction in adults and children.

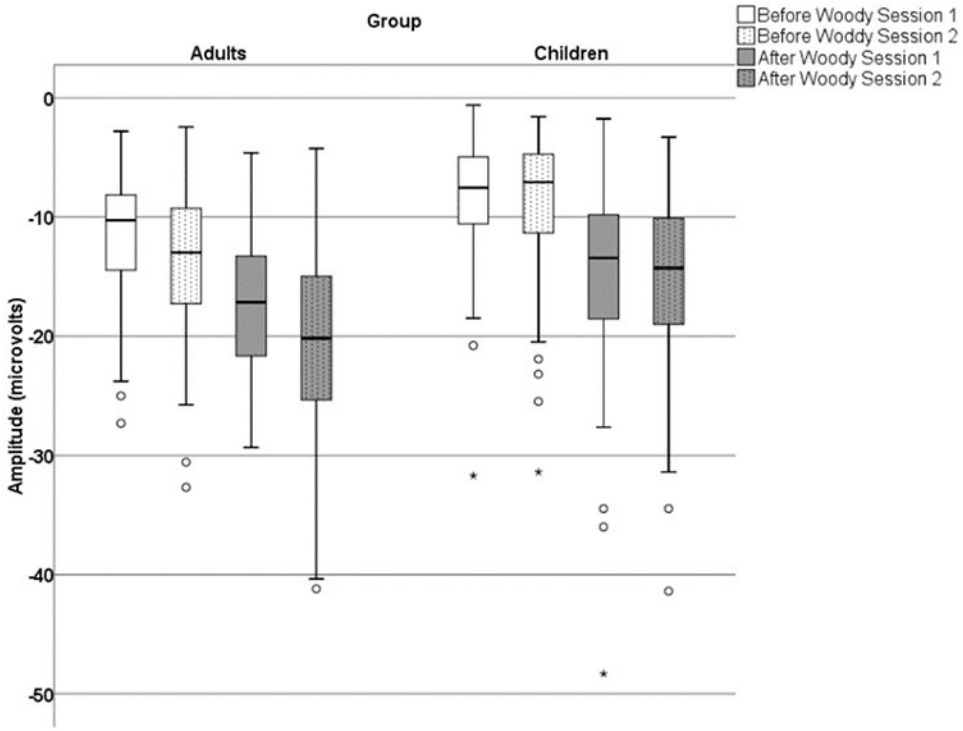

Figure 5.

The boxplot of ERN amplitude for session 1 and session 2 before and after latency jitter correction in adults and children.

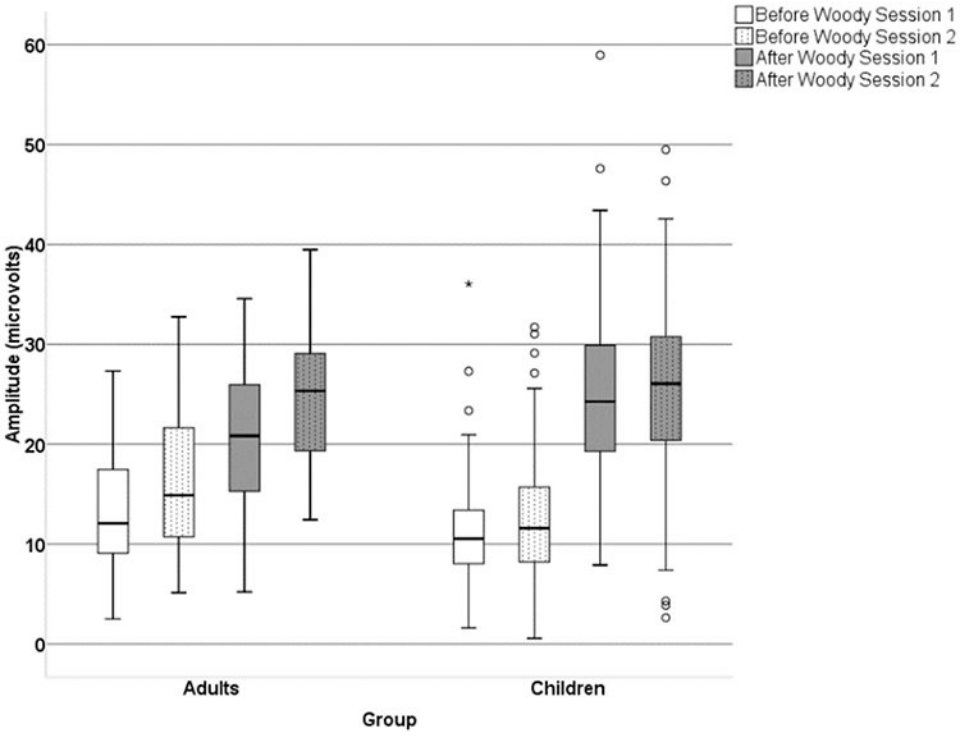

Figure 6.

The boxplot of Pe amplitude for session 1 and session 2 before and after latency jitter correction in adults and children.

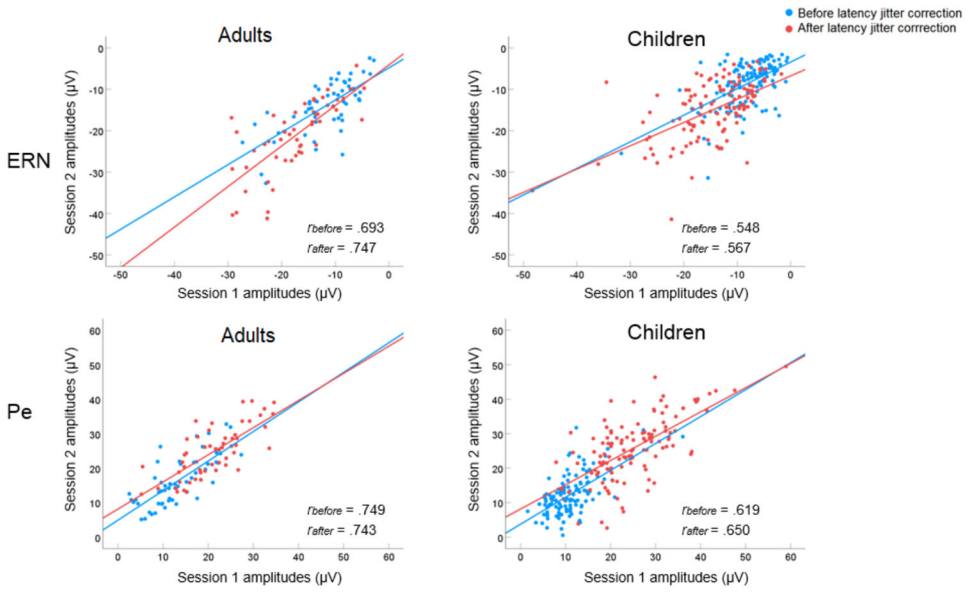

Figure 7.

The scatter plots depicting reliability of the ERN and Pe amplitudes between session 1 and session 2 before and after latency jitter correction in adults and children. Note: r_{before} represents the correlation coefficient between session 1 and session 2 before the latency jitter correction; r_{after} represents the correlation coefficient between Session 1 and Session 2 after the latency jitter correction.

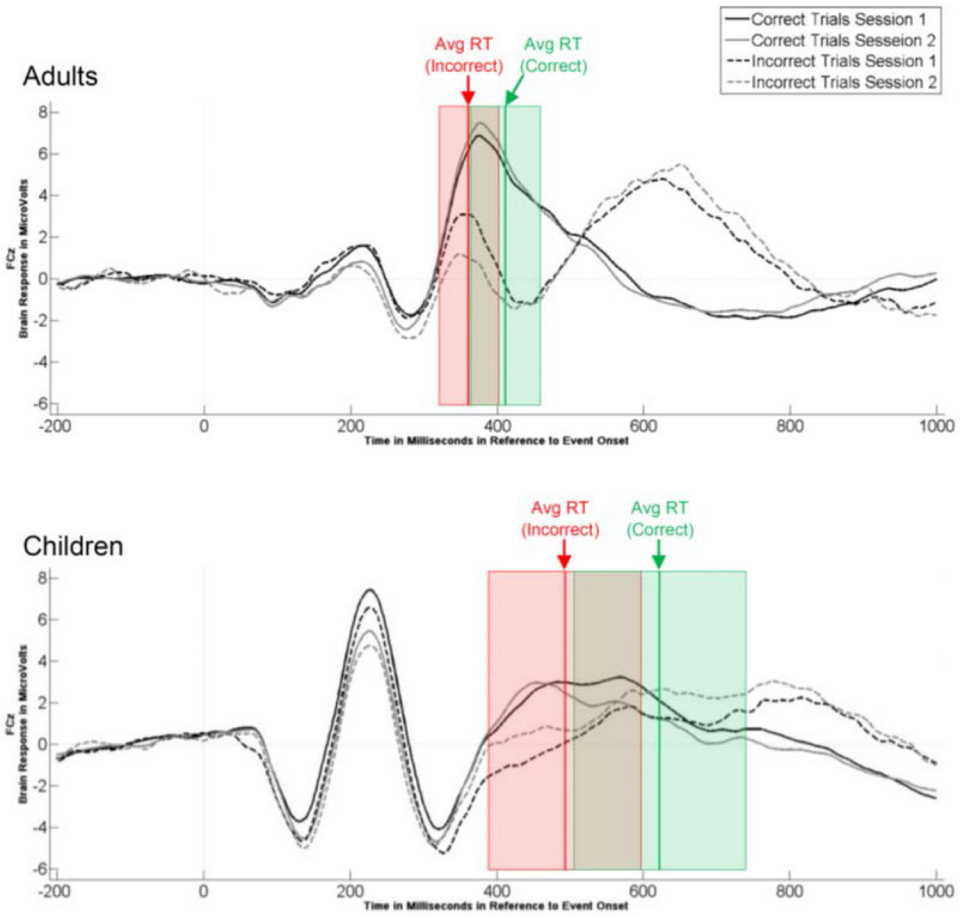

Figure 8.

Stimulus-locked ERPs for session 1 and session 2 for correct and incorrect trials in adults and children. Note: the dotted vertical red line represents the average reaction time on incorrect trials with the red colored box indicating +/− one standard deviation from this mean; dotted vertical green line represents the average reaction time on correct trials with the green colored box indicating +/− one standard deviation.

Table 1.

Participant distribution by age and sex after applying screening procedures and performance exclusion criteria.

Age Groups	Sex		Total
	Males	Females	
8	12	19	31
9	13	13	26
10	11	12	23
11	11	8	19
12	7	12	19
Adults	21	32	53
Total	75	96	171

Table 2.

A*verage number of segments (standard deviation) included in the averaged ERPs by age and session.*

	Sessions	
	Session 1	**Session 2**
Adults (n=53)	41.87 (20.24)	36.40 (19.65)
Children		
All (n=118)	65.77 (24.45)	46.81 (21.82)
8 yr (n=31)	62.58 (27.77)	51.55 (26.22)
9 yr (n=26)	77.27 (20.93)	52.35 (22.59)
10 yr (n=23)	56.83 (22.70)	40.83 (18.92)
11 yr (n=19)	65.16 (20.64)	43.37 (19.00)
12 yr (n=19)	66.68 (24.95)	42.16 (16.67)

Table 3.

Time windows for scoring stimulus-locked ERPs for adults and children

	P1 window (ms)	N1 window (ms)	P2 window (ms)	N2 window (ms)	P3 window (ms)
Adults	0-100	70-150	110-240	170-350	320-575
Children	0-100	70-170	130-270	200-375	320-600

Table 4.

Means, standard deviations of response times in correct and incorrect trials, and error rates for session 1 and session 2 in children and adults

	Sessions		Reliability		
	Session 1	Session 2	*r*	ICC (3,1) consistency	ICC (3,1) absolute agreement
Response times in correct trials (ms)					
Adults (n=53)	414.04 (45.98)	407.12 (47.94)	.81***	.81***	.80***
Children					
All (n=118)	623.56 (115.51)	620.38 (117.19)	.91***	.91***	.91***
8 yr (n=31)	720.39 (99.93)	699.24 (116.25)	.92***	.91***	.89***
9 yr (n=26)	662.19 (89.51)	664.23 (105.25)	.84**	.83***	.84***
10 yr (n=23)	586.72 (76.15)	589.64 (81.37)	.79***	.79***	.79***
11 yr (n=19)	594.95 (71.21)	595.08 (73.97)	.87***	.87***	.87***
12 yr (n=19)	485.90 (76.74)	494.23 (75.02)	.80***	.80***	.81***
Response times in incorrect trials (ms)					
Adults (n=53)	362.43 (40.71)	360.28 (39.93)	.77***	.77***	.77***
Children					
All (n=118)	478.93 (93.31)	506.47 (115.84)	.83***	.81***	.78***
8 yr (n=31)	542.96 (95.70)	570.03 (137.17)	.86***	.81***	.79***
9 yr (n=26)	504.95 (87.82)	530.41 (100.34)	.89***	.89***	.86***
10 yr (n=23)	440.60 (77.01)	483.82 (87.87)	.63***	.62**	.56**
11 yr (n=19)	474.10 (50.59)	491.56 (104.26)	.62***	.48*	.49*
12 yr (n=19)	390.09 (50.45)	412.34 (59.51)	.70***	.69***	.65***
Error rate (%)					
Adults (n=53)	9.07 (4.14)	7.70 (4.10)	.74***	.74***	.70***
Children					
All (n=118)	14.79 (5.21)	10.73 (4.92)	.62***	.62***	.47***
8 yr (n=31)	14.77 (5.77)	12.56 (6.16)	.77***	.77***	.72***
9 yr (n=26)	17.12 (4.42)	11.89 (4.69)	.63**	.63***	.38***
10 yr (n=23)	12.76 (4.85)	9.05 (3.90)	.43*	.42*	.32*
11 yr (n=19)	14.74 (4.93)	9.77 (4.04)	.56*	.55**	.35**
12 yr (n=19)	14.12 (5.20)	9.17 (3.70)	.51*	.48*	.31*

Note: the data were presented as mean (standard deviation); yr = year-old

*** $p < .001$

** $p < .01$

* $p < .05$

Table 5.

Means, standard deviations, and reliability indices of response-locked ERP component (ERN and Pe) amplitudes and latencies (ms) in Session 1 and Session 2 for incorrect trials before Woody filter

	Before Jitter Correction				
	Sessions		Reliability		
	Session 1	Session 2	*r*	ICC (3,1) consistency	ICC (3,1) absolute agreement
ERN amplitude (μV)					
Adults (n=53)	−11.80 (5.77)	−14.00 (6.49)	.69***	.69***	.65***
Children (n=118)	−8.10 (4.63)	−8.60 (5.43)	.55***	.54***	.54***
8 yr (n=31)	−7.92 (4.61)	−7.68 (4.76)	.48**	.48*	.49**
9 yr (n=26)	−6.11 (3.33)	−5.37 (2.06)	.11	.10	.10
10 yr (n=23)	−8.38 (4.31)	−8.56 (4.71)	.49*	.49*	.50**
11 yr (n=19)	−8.23 (2.54)	−9.78 (4.98)	.56*	.45*	.43*
12 yr (n=19)	−10.67 (6.80)	−13.39 (7.31)	.58*	.58*	.55**
ERN latency (ms)					
Adults (n=53)	68.08 (19.14)	67.94 (21.07)	.33*	.33**	.33**
Children (n=118)	46.36 (39.78)	44.66 (33.12)	.16	.16*	.16*
8 yr (n=31)	46.12 (56.87)	30.68 (29.21)	.03	.02	.02
9 yr (n=26)	31.17 (29.95)	38.80 (32.58)	.13	.13	.13
10 yr (n=23)	38.13 (19.24)	43.82 (33.33)	.04	.03	.04
11 yr (n=19)	62.45 (39.56)	50.32 (27.53)	.22	.21	.20
12 yr (n=19)	61.42 (26.43)	70.83 (31.25)	.22	.22	.21
Pe amplitude (μV)					
Adults (n=53)	13.20 (6.15)	16.23 (7.04)	.75***	.74***	.68***
Children (n=118)	11.19 (4.90)	12.48 (6.18)	.62***	.60***	.59***
8 yr (n=31)	11.53 (4.23)	10.74 (4.91)	.59***	.58***	.58***
9 yr (n=26)	9.16 (2.67)	10.93 (3.87)	.29	.28	.25
10 yr (n=23)	10.61 (3.84)	12.13 (6.66)	.41	.36*	.35*
11 yr (n=19)	11.70 (5.48)	14.17 (7.40)	.79***	.75***	.71***
12 yr (n=19)	13.61 (7.45)	16.17 (7.20)	.72**	.72***	.68***
Pe latency (ms)					
Adults (n=53)	187.28 (49.39)	195.04 (50.62)	.52***	.52***	.52***
Children (n=118)	149.27 (61.31)	151.67 (42.42)	.19*	.18*	.18*
8 yr (n=31)	132.37 (64.44)	130.20 (41.65)	.11	.10	.10
9 yr (n=26)	123.99 (38.69)	153.02 (31.30)	.18	.18	.13
10 yr (n=23)	163.17 (75.44)	156.25 (49.86)	−.05	−.04	−.05
11 yr (n=19)	171.57 (60.19)	160.31 (46.38)	.58*	.56**	.56**
12 yr (n=19)	172.34 (45.86)	170.64 (31.21)	−.25	−.24	−.25

Note: the data were presented as mean (standard deviation); the amplitude was calculated based on the peak-to-peak approach

*** $p < .001$

** $p < .01$

* $p < .05$

Table 6.

Means, standard deviations, and reliability indices of response-locked ERP component amplitudes (ERN and Pe) and latencies (ms) in Session 1 and Session 2 for incorrect trials after Woody filter

	After Jitter Correction				
	Sessions		Reliability		
	Session 1	Session 2	*r*	ICC (3,1) consistency	ICC (3,1) absolute agreement
ERN amplitude (μV)					
Adults (n=53)	−17.25 (6.57)	−20.99 (8.62)	.75***	.72***	.65***
Children (n=118)	−14.83 (7.05)	−15.08 (7.01)	.57***	.57***	.57***
8 yr (n=31)	−15.07 (7.74)	−13.98 (7.36)	.73***	.72***	.72***
9 yr (n=26)	−12.17 (3.92)	−11.19 (4.02)	.08	.08	.08
10 yr (n=23)	−13.84 (5.13)	−15.52 (6.29)	.59**	.58**	.57**
11 yr (n=19)	−15.07 (5.61)	−15.33 (4.69)	.49*	.49*	.50*
12 yr (n=19)	−19.00 (10.40)	−21.43 (8.32)	.38	.37	.37
ERN latency (ms)					
Adults (n=53)	70.59 (20.40)	67.01 (20.58)	.29*	.29*	.29*
Children (n=118)	47.12 (41.34)	48.32 (34.07)	.20*	.20*	.20*
8 yr (n=31)	41.65 (60.12)	32.64 (34.19)	.12	.10	.11
9 yr (n=26)	30.16 (22.30)	40.72 (33.42)	−.14	−.13	−.13
10 yr (n=23)	43.44 (23.95)	52.10 (30.49)	.05	.05	.05
11 yr (n=19)	65.12 (44.57)	51.55 (25.19)	.47*	.40*	.39*
12 yr (n=19)	65.69 (22.73)	76.53 (29.99)	.09	.09	.09
Pe amplitude (μV)					
Adults (n=53)	21.15 (7.03)	24.78 (7.44)	.74***	.74***	.66***
Children (n=118)	25.18 (8.40)	25.86 (9.10)	.65***	.65***	.65***
8 yr (n=31)	26.96 (8.75)	25.25 (9.04)	.69***	.69***	.69***
9 yr (n=26)	22.75 (4.57)	24.44 (6.56)	.53**	.50**	.49**
10 yr (n=23)	23.19 (8.45)	24.79 (10.36)	.53**	.52**	.52**
11 yr (n=19)	26.36 (8.23)	26.57 (8.89)	.71**	.71***	.72***
12 yr (n=19)	26.82 (11.12)	29.39 (10.63)	.74***	.74***	.73***
Pe latency (ms)					
Adults (n=53)	179.36 (43.10)	186.93 (45.84)	.53***	.53***	.52***
Children (n=118)	147.40 (48.26)	148.87 (39.54)	.18*	.18*	.18*
8 yr (n=31)	135.81 (58.96)	131.55 (38.01)	−.004	−.004	−.004
9 yr (n=26)	128.49 (30.52)	151.22 (29.80)	.27	.27	.22
10 yr (n=23)	154.81 (48.76)	148.40 (43.39)	−.01	−.01	−.01
11 yr (n=19)	171.26 (56.06)	154.91 (47.04)	.52*	.51*	.50*
12 yr (n=19)	159.39 (20.81)	168.43 (32.27)	−.13	−.12	−.12

Note: the data were presented as mean (standard deviation); the amplitude was calculated based on the peak-to-peak approach

*** $p < .001$

** $p < .01$

* $p < .05$

Table 7.

Means, standard deviations, and reliability indices of stimulus-locked ERP component amplitudes (N2, P2, N2, P3) and latencies (ms) in Session 1 and Session 2 for correct trials

	Correct Trials				
	Sessions		Reliability		
	Session 1	Session 2	*r*	ICC (3,1) consistency	ICC (3,1) absolute agreement
Adults					
N1 amplitude (μV)	−3.40 (1.74)	−3.30 (1.57)	.78***	.78***	.78***
N1 latency (ms)	111.90 (24.69)	114.15 (25.09)	.76***	.76***	.76***
P2 amplitude (μV)	7.03 (3.20)	6.58 (3.15)	.87***	.87***	.86***
P2 latency (ms)	190.52 (35.58)	186.27 (39.68)	.48***	.48***	.48***
N2 amplitude (μV)	−7.53 (2.94)	−7.69 (2.99)	.79***	.79***	.79***
N2 latency (ms)	276.02 (37.16)	269.24 (35.73)	.88***	.88***	.86***
P3 amplitude (μV)	11.40 (3.84)	12.69 (3.60)	.82***	.82***	.78***
P3 latency (ms)	381.73 (27.15)	380.69 (27.17)	.68***	.68***	.68***
P3 amplitude (μV) @Pz	11.71 (4.68)	12.52 (5.23)	.85***	.84***	.84***
P3 latency (ms) @Pz	382.43 (35.44)	371.65 (29.18)	.66***	.65***	.62***
Children					
N1 amplitude (μV)	−8.14 (2.85)	−8.67 (2.92)	.83***	.83***	.82***
N1 latency (ms)	127.97 (20.23)	132.18 (19.75)	.86***	.86***	.84***
P2 amplitude (μV)	16.51 (5.78)	15.02 (5.66)	.93***	.93***	.90***
P2 latency (ms)	223.15 (26.39)	221.89 (23.13)	.73***	.72***	.72***
N2 amplitude (μV)	−16.29 (6.89)	−14.82 (6.11)	.89***	.88***	.86***
N2 latency (ms)	321.98 (32.01)	317.94 (32.07)	.75***	.75***	.75***
P3 amplitude (μV)	11.84 (4.56)	12.50 (4.46)	.72***	.72***	.71***
P3 latency (ms)	447.30 (61.63)	442.99 (59.45)	.63***	.63***	.63***
P3 amplitude (μV) @Pz	16.50 (7.69)	15.71 (7.20)	.71***	.71***	.70***
P3 latency (ms) @Pz	421.79 (73.62)	407.62 (64.43)	.55***	.55***	.54***

Note: the data were presented as mean (standard deviation); the amplitude was calculated based on the peak-to-peak approach

*** $p < .001$

** $p < .01$

* $p < .05$

Table 8.

Means, standard deviations, and reliability indices of stimulus-locked ERP component amplitudes (N2, P2, N2, P3) and latencies (ms) in Session 1 and Session 2 for incorrect trials

	Incorrect Trials				
	Sessions		Reliability		
	Session 1	Session 2	*r*	ICC (3,1) consistency	ICC (3,1) absolute agreement
Adults					
N1 amplitude (μV)	−4.99 (2.69)	−5.06 (3.00)	.32*	.32*	.32*
N1 latency (ms)	114.39 (26.24)	114.29 (27.68)	.25	.25*	.25*
P2 amplitude (μV)	7.98 (3.13)	7.44 (3.71)	.63***	.62***	.62***
P2 latency (ms)	190.85 (36.48)	183.59 (38.20)	.40**	.40**	.40**
N2 amplitude (μV)	−9.31 (3.64)	−9.43 (4.13)	.74***	.73***	.74***
N2 latency (ms)	281.14 (34.57)	272.15 (37.22)	.75***	.74***	.72***
P3 amplitude (μV)	10.77 (3.91)	11.75 (4.83)	.51***	.50***	.49***
P3 latency (ms)	375.48 (58.57)	378.08 (56.24)	.58***	.57***	.58***
P3 amplitude (μV) @Pz	10.34 (4.71)	11.35 (5.22)	.59***	.58***	.58***
P3 latency (ms) @Pz	373.08 (56.72)	368.00 (55.42)	.69***	.69***	.69***
Children					
N1 amplitude (μV)	−9.87 (4.10)	−10.93 (4.67)	.37***	.37***	.36***
N1 latency (ms)	127.76 (21.46)	129.43 (23.53)	.59***	.58***	.58***
P2 amplitude (μV)	17.15 (6.57)	16.11 (6.20)	.64***	.64***	.63***
P2 latency (ms)	220.80 (28.01)	219.32 (25.99)	.58***	.57***	.58***
N2 amplitude (μV)	−17.29 (7.67)	−16.46 (6.65)	.75***	.74***	.74***
N2 latency (ms)	319.93 (32.04)	315.31 (42.21)	.40***	.38***	.38***
P3 amplitude (μV)	11.85 (5.20)	13.83 (5.51)	.58***	.58***	.54***
P3 latency (ms)	439.01 (65.28)	428.60 (58.62)	.41***	.41***	.41***
P3 amplitude (μV) @Pz	15.97 (7.12)	16.76 (8.46)	.62***	.61***	.61***
P3 latency (ms) @Pz	426.84 (72.90)	404.01 (76.52)	.35***	.35***	.33***

Note: the data were presented as mean (standard deviation); the amplitude was calculated based on the peak-to-peak approach

*** $p < .001$

** $p < .01$

* $p < .05$