Psychophysiology, 50 (2013), 111–123. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2012.01476.x

A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring

DOROTHEA HÄMMERER,[a,b] SHU-CHEN LI,[a,b] MANUEL VÖLKLE,[a] VIKTOR MÜLLER,[a] AND ULMAN LINDENBERGER[a]

[a]Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
[b]Institut für Pädagogische Psychologie und Entwicklungspsychologie, Technische Universität Dresden, Dresden, Germany

Abstract

The reliability, stability, and signal-to-noise ratio (SNR) of event-related potentials (ERPs) were investigated in children, adolescents, younger adults, and older adults in performance monitoring tasks. P2, N2, P3, and P2-N2 peak-to-peak amplitude showed high odd-even split reliabilities in all age groups, ranging from .70 to .90. Multigroup analyses showed that test-retest stabilities (across 2 weeks) of ERP amplitudes did not differ among the four age groups. In contrast, relative to adolescents and younger adults, SNRs were lower in children and older adults, with higher noise levels in children and lower signal power in older adults. We conclude that age differences in the SNR of stimulus-locked ERPs can be successfully compensated by the averaging procedure with about 40 trials in the average. However, age differences in baseline noise and split-half reliability should be considered when comparing age groups in single trial measures or time-varying processes with ERPs.

Descriptors: Reliability, ERP, Development, Aging, Performance monitoring

Event-related potentials (ERPs) offer noninvasive and temporally precise indicators for studies on cognitive development across the lifespan by providing insights into cortical processes that contribute to observed age-related differences in behavior. Knowledge about the reliability, stability, and signal-to-noise ratio (SNR) of various ERP components helps to judge how much of the variance represented in a given ERP is informative with regard to the task-relevant brain mechanisms that the ERP is supposed to capture. If researchers wish to interpret age group differences in ERPs, information about age-related differences in their measurement properties of ERPs is crucial for drawing meaningful developmental inferences (e.g., Labouvie, 1980).

However, despite a growing number of studies investigating age differences in ERPs, a comparison of the reliability, test-retest stability, and SNR of ERPs in multiple age groups spanning the lifespan is lacking. Most studies investigating different age groups do not explicitly test the comparability of reliability and stability measures in different age groups. Evidence on the reliability or stability of ERPs in different age groups is quite diverse, with test-retest stabilities and odd-even split reliabilities ranging from .10 to .80 (child development: Joutsiniemi et al., 1998; Segalowitz & Barnes, 1993; Segalowitz et al., 2010; Uwer & von Suchodoletz, 2000; adult development: Joutsiniemi et al., 1998; Sandman & Patterson, 2000; Thesen & Murphy, 2002; Walhovd & Fjell, 2002). This considerable variation may in part reflect differences between studies in artifact preprocessing, in the numbers of trials that constituted the ERP, or both. Also, it is to be expected that the number, location, and strength of the generating fields reflected in a given ERP, that correspond to different cognitive processes, affect its test-retest reliability. Clearly, comparisons of different age groups on the same tasks within one study are needed to answer the question whether the measurement properties of ERPs are comparable across the lifespan. A recent study investigated the number of trials necessary for achieving a reliable estimate of response-locked ERPs across the lifespan within one study (Pontifex et al., 2010). Reliable estimates were observed in children, younger adults, and older adults with about six trials in the average. This result is encouraging and suggests comparable measurement properties of response-locked ERPs across the lifespan. However, Pontifex et al. (2010) did not test directly for age group differences in ERP reliability.

Furthermore, one may expect that signal and noise levels of ERPs should also differ across age, reflecting differences in developmentally malleable features such as neuromodulation or brain structure, which may influence the level and dynamics of background noise in neural information processing (for reviews,

This study was supported by a German Research Foundation grant to Shu-Chen Li, Hauke Heekeren, and Ulman Lindenberger for a subproject (Li 515/8) in the research group on Conflicts as Signals (DFG FOR 778). We thank our student research assistants and interns Kirsten Becker, Angelika Paul, Katja Breitenbach, Beate Czerwon, Minh Tam Luong, Carlos Picchioni, Viola Störmer, Natalie Trumpp, and Katja Zschenderlein for their valuable support during data collection.

Address correspondence to: Dorothea Hämmerer, TU Dresden, Zellescher Weg 17, Fakultät Mathematik und Naturwissenschaften Fachrichtung Psychologie, 01062 Dresden, Germany. E-mail: haemmerer@mpib-berlin.mpg.de

PSYCHOPHYSIOLOGY

see Li, von Oertzen, & Lindenberger, 2006; MacDonald, Nyberg, & Bäckman, 2006; McIntosh et al., 2010; Winterer & Weinberger, 2004). It is unclear whether age differences in noise (assessed as variability in a baseline period, see Method for details) relate to age differences in the stability and reliability of ERPs or whether the averaging procedure can compensate successfully for the expected age differences in noise. Also, other electroencephalogram (EEG) measures such as dipole analyses or single trial analyses are more prone to age differences in SNR. Hence, it is important to quantify age differences in SNR even if the reliability and stability of ERPs are relatively unaffected by such differences.

In this study, we compare the odd-even split reliability and test-retest stability of stimulus-locked ERPs in four age groups covering the age periods from middle childhood to old age. In addition, the SNR of ERPs across the lifespan is assessed to investigate whether lifespan differences in reliability and stability are related to age differences in SNR. Since the number of trials included in the ERP is an important factor for the reliability and SNR of ERPs and might be a way to attain comparable measurement properties across the lifespan, we additionally compare reliability and SNR with increasing number of trials contributing to the average across the four age groups.

Sources of Variability in ERPs and Measures of Reliability, Stability, and SNR

Several sources of variability in ERPs can be distinguished: (a) variance related to measurement error, (b) individual changes in true score across measurement sessions, and (c) group changes in true score across sessions, such as training effects across sessions. While the first type of variability affects the reliability of the acquired measure, variability according to (b) and (c) will affect the test-retest stability of the measure (cf. Brim & Kagan, 1980). As outlined by Segalowitz and Barnes (1993), these different types of variability in ERP measures can be assessed, respectively, by comparing (a) the odd-even split reliability, (b) the Pearson product moment test-retest correlation, and (c) the absolute agreement intraclass test-retest correlation coefficient (ICC). Accordingly, the present study makes use of all three measures to compare the measurement properties of ERPs across the lifespan.

To further inform the underlying causes for potential age differences in reliability or test-retest stability, two additional analyses were performed: (1) odd-even and split-half reliability were compared to estimate the relative contributions of noisy measurements and gradual shifts in the course of a session. If split-half correlation coefficients were lower than odd-even split correlation coefficients, this would suggest that changes in ERP amplitude during the course of the recording session lowers overall within-session reliability; (2) measures of SNR were assessed to explore whether age-related differences in the reliability and stability of ERPs, if observed, are likely to reflect age differences in baseline noise levels, signal strength, or both.

ERPs Related to Performance Monitoring Investigated in the Present Study

Performance monitoring refers to a class of processes that come into play in challenging and nonroutine situations that involve response conflict or undesired action outcomes (see Botvinick, Braver, Barch, Carter, & Cohen, 2001; Holroyd & Coles, 2002). The ability to monitor response conflict varies considerably by age (e.g., Li, Hämmerer, Müller, Hommel, & Lindenberger, 2009). The ERP components investigated here include the P2, N2, and P3 component, as well as the P2-N2 peak-to-peak component. We thus investigate electrophysiological correlates that are relevant indicators of age differences in underlying cognitive functions. Furthermore, a broad range of cognitive functions is covered, including attentional orientation or attentional updating, action monitoring, and motor control (see below). Finally, to add to the generalizability of our findings, the ERPs are assessed in two tasks from different monitoring domains: a response conflict task and a reinforcement learning task.

A P2 can be observed in a time window of 100 to 250 ms following a visual stimulus indicating a critical event such as an imperative stimulus or a performance feedback (Hämmerer, Li, Müller, & Lindenberger, 2011; Jonkman, 2006). About 200 to 300 ms after the imperative stimulus or the feedback stimulus, a negative deflection (i.e., N2) can be noted. The N2 is larger the more response or outcome conflict is experienced (Holroyd, Hajcak, & Larsen, 2006; Miltner, Braun, & Coles, 1997). Following the N2, about 300 to 500 ms after the imperative stimulus or feedback stimulus, a positive deflection, referred to as P3, is observed. In a response inhibition task, the P3 after NoGo stimuli (NoGo-P3) has been linked consistently to withholding a response (Bekker, Kenemans, & Verbaten, 2004). The NoGo-P3 is larger than the P3 after Go stimuli (Go-P3) and shows a central scalp distribution, whereas the Go-P3 is strongest at parietal electrodes (Bruin & Wijers, 2002). In a reinforcement learning task, the P3 following a feedback has been related to the expectedness of this feedback for a specific response, being larger when the feedback is less expected (Campbell, Courchesne, Picton, & Squires, 1979).

Aims of the Study

In light of scarce and inconsistent age-comparative evidence on the reliability and stability of ERPs, the first aim of this study was to explore the extent to which odd-even split reliabilities and test-retest stabilities of ERP components assessed during performance monitoring would differ across the age groups. Unlike previous studies, we explicitly tested for age differences in reliability and stability using multigroup comparisons that test whether correlations observed in different age groups differ significantly from each other. Furthermore, given that the reliability of the ERP depends on the number of trials available for the average, we assessed whether different numbers of trials would be needed across the lifespan to reach satisfactory and comparable levels of within-session reliability.

Second, we compared odd-even split and split-half reliability on ERP measures across age groups to gauge age group differences in the amount of intraindividual variability within a testing session. Given that intraindividual performance variability at the behavioral level tends to be larger in children and older adults (cf. Li, Lindenberger, Hommel, Aschersleben, Prinz, & Baltes, 2004; Lövdén, Li, Shing, & Lindenberger, 2007; MacDonald et al., 2006; Papenberg et al., 2011), we expected that split-half correlations would be reduced in these age groups.

Third, we investigated whether the SNR of ERPs varies across age groups. Signals are expected to be stronger in age groups with thinner skulls, that is, in infants, children, and adolescents (cf. Knott, Hazony, Karafa, & Koltai, 2004; Lamm, Zelazo, & Lewis, 2006). At the same time, noise levels are assumed to be larger in children and older adults (cf. Lövdén et al., 2007, McIntosh et al., 2010). Altogether, then, one would expect lower SNR in children and older adults. Taken together, delineating the pattern of age

group differences in reliability, stability, and SNR is of methodological and substantive interest for age-comparative investigations of ERP signals at various levels of aggregation.

Method

Participants

The study sample included a total of 185 participants covering four age groups: 45 children (22 girls, mean age = 10.15 years, *SD* = 0.60), 46 adolescents (22 women, mean age = 14.38 years, *SD* = 0.55), 47 younger adults (22 women, mean age = 24.27 years, *SD* = 2.07), and 47 older adults (24 women, mean age = 71.24 years, *SD* = 2.91). Participants were invited for two sessions that were 2 weeks apart (mean: 14.32 days, *SD* = 2.59 days).

The data of several participants had to be excluded from the analysis because of technical problems during EEG recording, or because participants did not reach the minimum learning criteria or did not comply with the task instructions. The present study comprised two tasks that were administered in two testing sessions each. In each session, more than 60 trials were on average included for each age group and condition.

Our sample was typical with respect to lifespan changes in perceptual speed (planned curvilinear contrast: $t = 11.6$, $p < .05$, $d = 1.84$), and verbal knowledge (increase with increasing age: $\chi^2(3, N = 184) = 129.8$, $p < .05$).

Experimental Procedure

During EEG recordings, participants were comfortably seated in an electrically and acoustically shielded room. The distance to the computer screen was 80 cm. In both sessions, the participants first worked on a probabilistic reinforcement learning task and then on a speeded version of the cued Continuous Performance Task (CPT).

During the reinforcement learning task, participants were presented with pairs of Japanese characters that were each associated with probabilistic gains and losses. However, within each pair, one symbol had a higher probability of leading to a gain than the other symbol. Subjects were asked to maximize gains by identifying the option with a greater gain probability in each pair (cf. Frank, Seeberger, & O'Reilly, 2004; for further details on the task procedure, see Hämmerer et al., 2011). The task used in the second testing session was identical to the one of the first session, except that it included three new sets of the three pair types in which the better symbol had to be identified.

In the second part of each testing session, a modified version of the cued CPT was administered (Braver et al., 2001; Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956). The task was adapted to be suitable for testing children by replacing the letter stimuli with color stimuli. Participants were instructed to respond by pressing a button with their right index finger as fast as possible when the blue square was followed by the yellow square. The critical NoGo condition is a cue stimulus followed by a nontarget (e.g., blue square followed by a red square). The task used in the second testing session was identical to the one of the first session, except that it employed a new sequence of pair presentations to exclude the possibility that implicit sequence learning might influence the performance in the second session.

EEG Recordings and Data Preparation

EEG was recorded from 64 Ag/AgCl electrodes placed according to the 10-10 system in an elastic cap (BrainCap, Brain Products GmbH), using BrainVision Recorder. The sampling rate was 1000 Hz with a band-pass filter applied in the range of 0.01 to 250 Hz. EEG recordings were referenced online to the right mastoid. The ground was positioned above the forehead. Impedances were kept below 5 kΩ.

Using BrainVision Analyzer, the recorded data were rereferenced to an average reference. Using the FieldTrip software package (for more details, see http://www.ru.nl/fcdonders/fieldtrip), the data were then segmented into epochs of 2 s before and 2.5 s after the onset of the colored square (CPT) or feedback (reinforcement learning task). Epochs or channels with severe muscular artifacts or saturated recordings were excluded manually. In the CPT, an average of 12% of the trials in the first session and 16% in the second session had to be excluded (children: 19% [24%], adolescents: 13% [19%], younger adults: 8% [11%], older adults: 8% [10%]; data for the second session in brackets). In the reinforcement learning task, an average of 7% of the trials in the first session and 9% in the second session had to be excluded (children: 9% [11%], adolescents: 6% [10%], younger adults: 8% [8%], older adults: 6% [6%]; data for the second session in brackets).

For both tasks, the thus preprocessed data were further subjected to an independent component analysis (ICA) decomposition using EEGLAB (Delorme & Makeig, 2004) for artifact rejection. ICA components of ocular and muscular artifacts were removed from the data. The recombined data were band-pass filtered in the range of 0.5 to 25 Hz and epoched 1,000 ms after and 100 ms before the onset of the feedback symbols. Baseline corrections were applied on the epoched data with respect to the 100-ms prestimulus baseline. ERPs were obtained by averaging across trials for each electrode and condition for each participant. Amplitudes of the P2, N2, and P3s following the feedback or a colored square were defined as the most positive (or negative) peaks in the individual averages in the time windows 100 to 250 ms, 200 to 350 ms, and 300 to 500 ms, respectively. The amplitude difference between the P2 and N2 peaks was examined in addition to the absolute N2 peak to also incorporate the possibility that the N2 might be superimposed on a positive deflection (cf. Hämmerer et al., 2011). We focused on peak instead of mean area measures of a specified time window since a comparison of mean measures across different age groups might be biased by age differences in the slope of the ERP (e.g., Jonkman, Lansbergen, & Stauder, 2003). ERP plots of the age groups and sessions can be seen in Figures 1 and 2. Also, adult age differences in the slope of ERPs have been shown to be independent of age differences in temporal jitter of the single ERPs (Walhovd, Rosquist, & Fjell, 2008).

Finally, in addition to an individual assessment of the ERPs to critical and noncritical events (i.e., loss and gain feedback) during outcome monitoring, the reliability of their difference is assessed to also examine the reliability of ERP measures reflecting conflict cost.

Statistical Analyses

Data were analyzed using SPSS (version 15), AMOS 16.0.1, and SAS (SAS 9.1.3, Windows version 5.2.3790). As previous studies showed age differences in EEG scalp distributions across the life span (e.g., Müller, Brehmer, von Oertzen, Li, & Lindenberger, 2008), we analyzed the data at the single electrode level rather than clustering the electrodes. To identify the electrodes with the largest effects in each age group and condition, multivariate analyses of variance (MANOVAs) were performed for each ERP and experimental condition on 35 leads (AF7, Fp1, Fpz, Fp2, AF8, F7, F3, Fz,

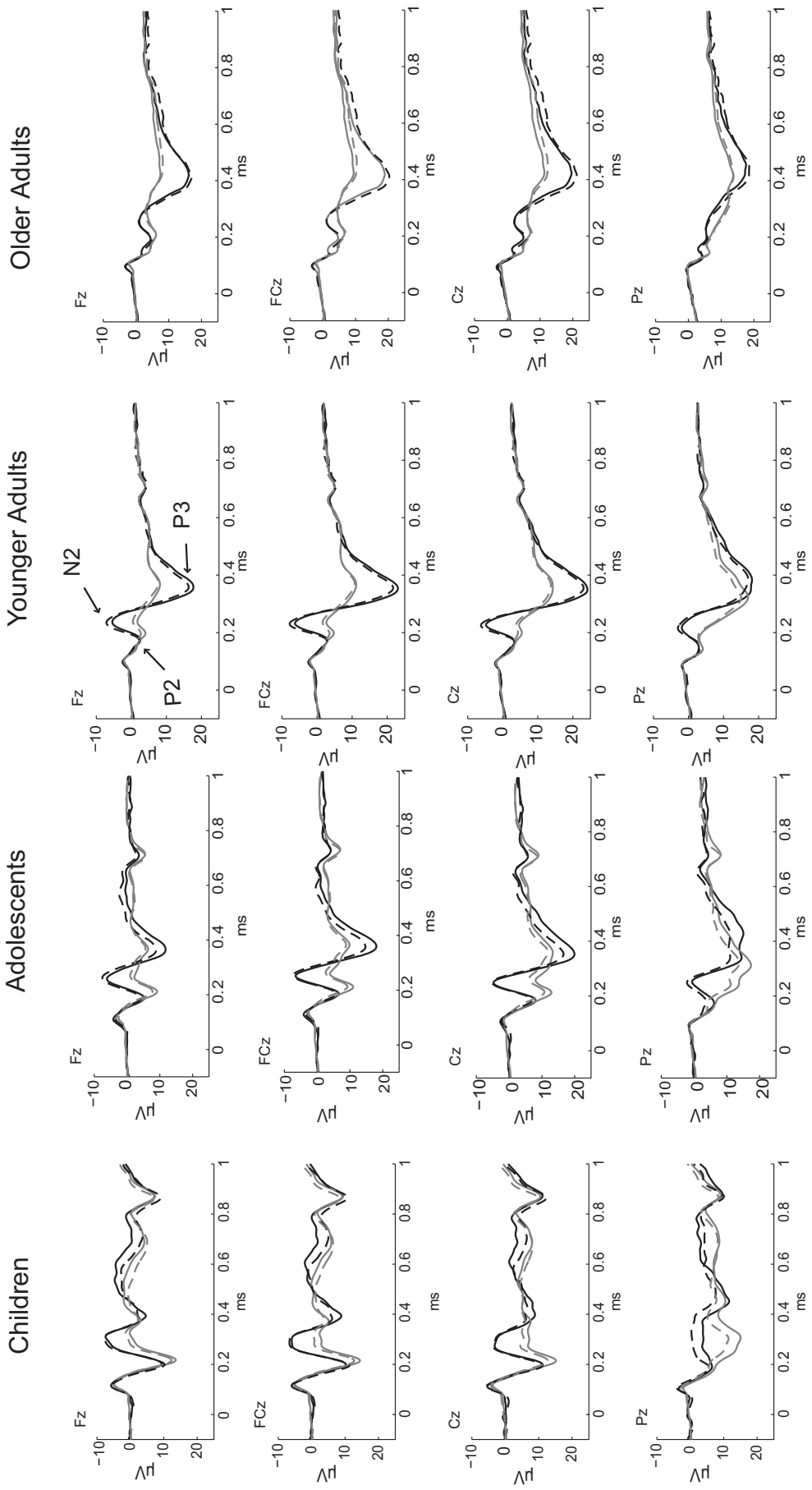

Figure 1. Grand average of the stimulus-locked ERPs in Go and NoGo conditions across the four age groups. Four midline electrodes (Fz, FCz, Cz, and Pz) are displayed. P2, N2, and P3 ERP components are indicated by arrows.

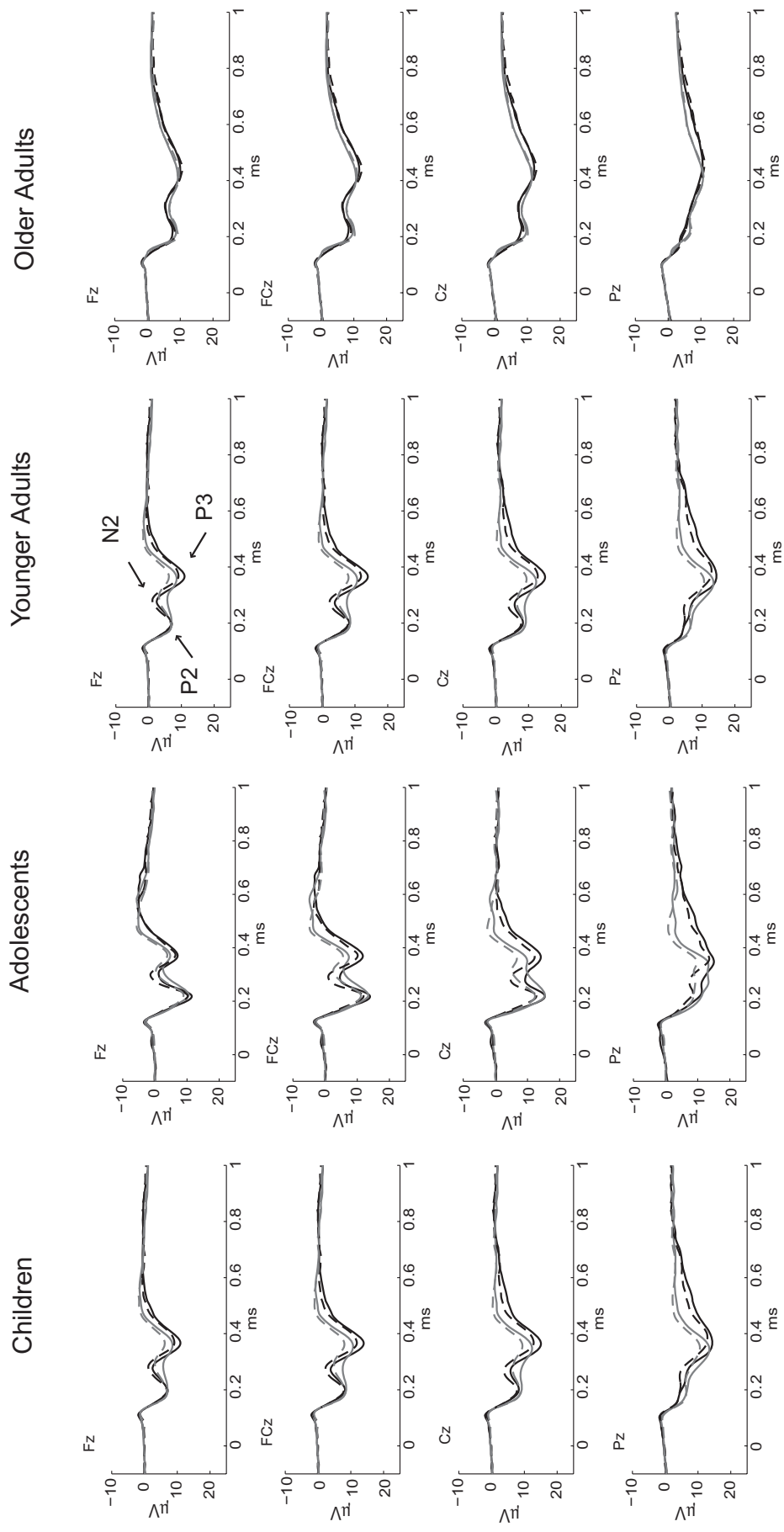

Figure 2. Grand average of the stimulus-locked ERPs in gain and loss conditions across the four age groups. Four midline electrodes (Fz, FCz, Cz, and Pz) are displayed. P2, N2, and P3 ERP components are indicated by arrows.

Table 1. *Electrodes with Strongest Effects for ERP Components Under Investigation*

	CPT task Go trials				Reinforcement learning task Gain trials				Reinforcement learning task Loss trials			
	P2	N2	P3	P2-N2	P2	N2	P3	P2-N2	P2	N2	P3	P2-N2
Children	Cz (FCz)	Fpz (Fpz)	Pz (Pz)	Fz (Fz)	Cz (FCz)	Fp1 (Fpz)	P4 (P4)	Fz (Fz)	FCz (FCz)	Fp1 (Fz)	Pz (Pz)	Fz (Fz)
Adolescents	Cz (CPz)	Fp1 (Fp1)	Pz (CPz)	Fz (Fz)	Cz (CPz)	Fp1 (Fp1)	P4 (P4)	FCz (FCz)	Cz (Cz)	Fp1 (Fp1)	Pz (CPz)	Fz (Fz)
Younger adults	Cz (CPz)	Fp1 (Fpz)	Pz (CPz)	Fz (Fz)	Cz (CPz)	Fp1 (Fpz)	CP4 (P4)	Fz (Fz)	Cz (Cz)	Fpz (Fp1)	Pz (CPz)	FCz (FCz)
Older adults	Cz (Cz)	Fpz (Fpz)	CPz (CPz)	Fz (Fz)	FCz (FCz)	Fpz (Fpz)	CPz (Cz)	Fz (FCz)	FCz (FCz)	Fp2 (Fp2)	CPz (CPz)	Fz (Fz)

Note. Electrodes for the first session are reported without brackets, and electrodes for the second session are reported in brackets.

F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, PO7, O1, Oz, O2, PO8), including age group as the between-subjects factor (children, adolescents, younger adults, and older adults) and laterality (5 levels: left, medium-left, midsagittal, medium-right, right), as well as anterior-posterior (7 levels: frontopolar, frontal, frontocentral, central, centroparietal, parietal, occipital) as within-subjects factors. Given that scaling methods have the potential to distort the nature of group differences if groups differ in within-groups variability (which is the case in age group comparisons), analyses were performed on unscaled data (cf. Haig, Gordon, & Hook, 1997). As can be seen in Table 1, peak electrode sites differed only marginally across the age groups. The P2 was maximal at central electrodes, the N2 was maximal at frontal electrodes, and the P3 was maximal at centroparietal electrodes. These localizations are in line with the scalp distributions observed in prior developmental studies (e.g., Falkenstein, Hoorman, & Hohnsbein, 2002; Jonkman, 2006). The peak data of the odd-even split data were chosen from the electrodes with the largest effects in the unsplit data, assuming that the localization of the effect is more precise when all trials are included.

Reliability and Stability Measures

Variables that differ in variance between age groups and are associated with the ERPs may contribute to age group differences in reliability. Hence, the EEG data from the first and second session were residualized within age groups with respect to the number of trials that entered the average ERPs, and to differences in task performance (i.e., percentage of correct choices for each pair in the reinforcement learning task and percentage of prime-based NoGo errors as well as median reaction time in the CPT, cf. Lamm et al., 2006). Thereafter, odd-even split, test-retest, and intraclass correlations were controlled within age group for individual differences in age and perceptual speed (digit symbol substitution test). Odd-even split correlations were adjusted for test length using the Spearman-Brown formula. The ICCs chosen were absolute agreement ICCs for a two-way mixed model (persons random, sessions fixed) for single measures. Absolute agreement ICCs were chosen to include variance across sessions on the group level.

Multigroup analyses on data z standardized within each age group were performed for each ERP to assess age differences in test-retest correlations in ERP amplitude. The EEG data from the first session were used to predict the EEG data from the second session. Corresponding to the correlational analyses for the separate age groups, residualized EEG measures were entered into the analyses. In general, the variances of the EEG data did not differ reliably between session 1 and session 2 within age groups; only in 7 out of 112 reliability tests were significant session differences within age groups observed. To test whether test-retest stabilities differed between age groups, an unconstrained model allowing all correlations to vary between age groups was compared to a model where the correlations were constrained to be the same for all age groups. Regressions predicting the ERP measures from the second session by the ERPs from the first session were calculated for all four age groups to identify influential individual cases. On average, per regression, 0.875 of the approximately 170 cases tested for each ERP were removed (Cook's distance > .20).

Comparison of Odd-Even and Split-Half Correlations as a Function of Increasing Number of Trials in the Individual Averages

Individual averages for the odd-even split and split-half correlations were calculated on an increasing number of randomly selected trials (from 5 up to 100 trials, in steps of 5 trials) for the Go-P3 amplitude. The Go-P3 was chosen because it is a well-established ERP component that can be observed on a large variety of tasks that include response execution (e.g., Bruin & Wijers, 2002). Focusing on the P3 is hence of relevance for a broad range of age-comparative EEG paradigms. Odd-even split correlations were calculated by randomly selecting an increasing number of trials from the odd and the even trials of a given session. Split-half correlations were calculated by choosing trials randomly from the first and the second half of the session. To achieve a more reliable estimate of the correlations of randomly selected trials, we report the mean of 200 random trial selections. Individual averages were based on the mean in a 20-ms window around the maximum peak of the unsplit Go-P3 in each subject, as individual averages on very few trials were too noisy for a peak measure approach.

SNR

Noise measures were assessed during the 100-ms baseline interval, where variability in EEG measures should be unrelated to variability in the strength of the stimulus-locked response (cf. e.g., Maidhof, Rieger, Prinz, & Koelsch, 2008). For the SNR, the square root of signal power (mean of the squared signal in a 20-ms time window around individually defined peaks) was divided by the square root of baseline noise (variance in the baseline interval on the average of the selected trials). As for the reliability analyses above, this process was repeated 200 times. Repeated measures

MANOVAs that allow for age differences in variance with the factors age groups and numbers of trials were conducted. Reliable age differences were followed up with planned contrasts.

Results

Odd-Even Split and Split-Half Reliabilities as a Function of Trial Number

As can be seen in Panel A of Figure 3, odd-even split correlation coefficients increased steadily with an increasing number of trials in the average measure, reaching an asymptote at approximately 40 trials. Indicating highest noise levels in children, odd-even split correlation coefficients were consistently lower in the children-age than in the other three age groups. Further, when comparing the split-half and odd-even split correlation coefficients in Panel A of Figure 3, it can be seen that the split-half correlation coefficients reached a somewhat lower asymptote than the odd-even split correlations in children, adolescents, and older adults (split-half correlation coefficients: children .81, adolescents .91, younger adults, .96, older adults .89; odd-even split correlations: children .91, adolescents .96, younger adults .97, and older adults .96 with 100 trials in the average). This suggests that interindividual differences in the change of the Go-P3 amplitude within a recording session were smaller in younger adults than in the other age groups. Interestingly, as can be seen in Panel B of Figure 3, the mean amplitude of the Go-P3 decreased considerably in children during the recording session of about 30 mins, but stayed at a similar level in older adults. At the same time, not only children but also older adults showed lower split-half than odd-even split reliabilities. This suggests that the group of older adults showed interindividual differences in the slope of ERPs during a recording session in the presence of stable group average ERP amplitudes.

Odd-Even Split Reliabilities and Test-Retest Stabilities

The ERPs for both sessions and tasks are shown in Figures 1 and 2. Overall, the odd-even split correlations of the amplitudes were rather high, indicating satisfactory measurement reliability in all age groups. All correlation coefficients of the conditions were higher than .70, with most of them being higher than .80.

Pearson product moment correlations and intraclass correlations were calculated for each ERP and age group. As can be seen in Table 2, the Pearson product moment correlation coefficients differed only marginally from the intraclass correlation coefficients. This suggests marginal changes in true score on the group level across sessions.[1] The P2 and P3 showed acceptable stabilities with ICCs above .48 in all four age groups (see Figure 4). The N2 in the go, gain, and loss condition in children was somewhat lower, ranging from .38 to .50.

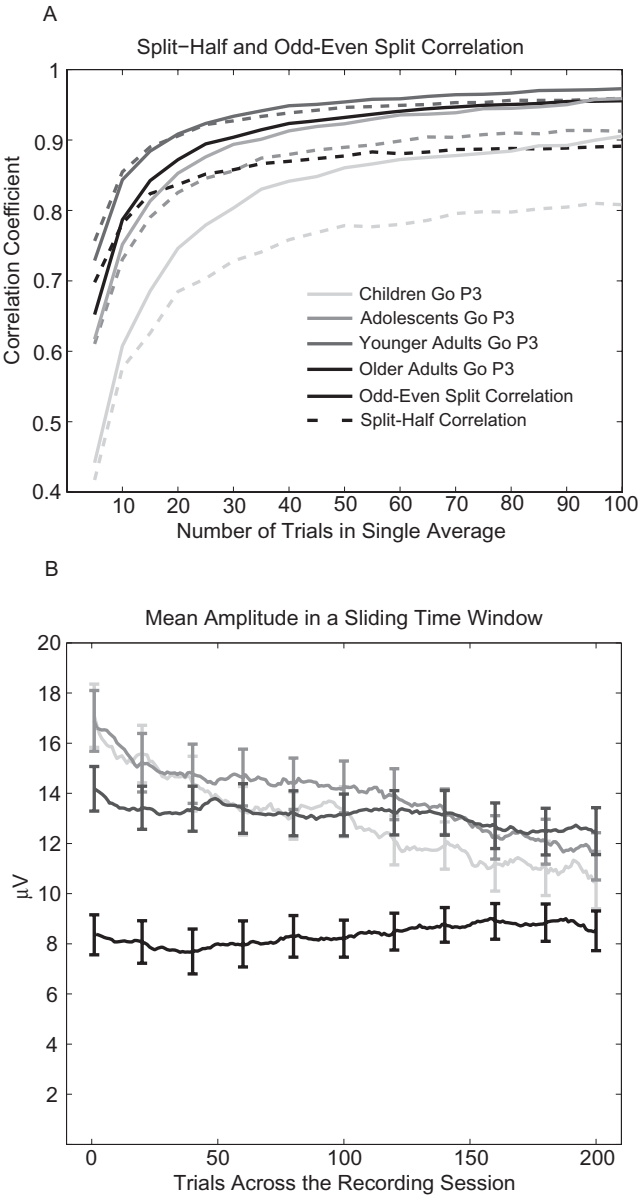

Figure 3. A: Odd-even split and split-half correlations of the Go-P3 amplitude with increasing trial number in the single averages. Correlations are based on the average in a 20-ms time window around the individual peaks of unsplit single averages. The Spearman-Brown equation was used to adjust the correlation coefficients for test length. Error bars denote 1 *SE* (according to Cleary & Linn, 1969). B: Change of Go-P3 amplitude during the EEG recording across a duration of 200 trials (approximately 30 min). Single averages are calculated within a trialwise sliding time window of 40 trials. The mean of these single averages per age group is depicted. Error bars denote 1 *SE*.

Multigroup Analyses Assessing Age-Group Differences in Test-Retest Stabilities

To test for age differences in test-retest stabilities, multigroup analyses were performed. These compared an unconstrained model to a model in which the test-retest correlations were assumed to be the same across the four age groups. As can be seen in Table 3, the test-retest stabilities of the ERP amplitudes in children, adolescents, younger adults, and older adults did not differ significantly.

1. By design, there are too few NoGo trials for reliability analyses. For completeness of the reported results and also to indicate the impact of small trial numbers on the test-retest stability of the ERP measure, the ICC coefficients for the NoGo condition shall nonetheless be given here: For children, adolescents, younger adults, and older adults, the ICCs for the P2 amplitude were .19 (*n.s.*), .47, .69, and .89, respectively; for the N2 amplitude, they were .12 (*n.s.*), .56, .54, .76; for the P3 amplitude, they were .29, .59, .70, .78; for the P2-N2 peak difference, they were .28, .49, .67, .73. If not indicated otherwise, all correlations reported in this footnote were reliable at the alpha level of $p < .05$.

Table 2. *Pearson Product Moment and Intraclass Test-Retest Correlations for the ERP Amplitudes Across the Two Testing Sessions*

	CPT task Go trials				Reinforcement learning task Gain trials				Reinforcement learning task Loss trials				Reinforcement learning task Amplitude loss-gain			
	P2 Go	N2 Go	P3 Go	P2-N2 Go	P2 Gain	N2 Gain	P3 Gain	P2-N2 Gain	P2 Loss	N2 Loss	P3 Loss	P2-N2 Loss	P2 Loss-gain	N2 Loss-gain	P3 Loss-gain	P2-N2 Loss-gain
Children	.72*	.41*	.62*	.65*	.68*[1]	.40*	.65*	.59*[1]	.52*[1]	.51*	.54*	.52*[1]	.00	.20	.36*	.42*
	(.72*)	(.40*)	(.61*)	(.65*)	(.67*[1])	(.37*)	(.66*)	(.58*[1])	(.54*[1])	(.50*)	(.53*)	(.50*[1])	(-.04)	(.18)	(.31*)	(.43*)
Adolescents	.69*	.62*	.61*	.68*[1]	.78*	.66*	.68*	.66*	.78*	.64*	.60*	.66*	.24	.39*	.46*	.44*
	(.66*)	(.59*)	(.60*)	(.69*[1])	(.76*)	(.68*)	(.69*)	(.64*)	(.77*)	(.66*)	(.59*)	(.68*)	(.18)	(.35*)	(.47*)	(.43*)
Younger adults	.85*	.80*	.74*[1]	.76*	.71*	.61*	.69*	.71*	.49*	.67*[1]	.63*	.63*	.26*	.37*	.59*	.41*
	(.83*)	(.79*)	(.56*[1])	(.76*)	(.70*)	(.61*)	(.65*)	(.72*)	(.48*)	(.65*[1])	(.62*)	(.62*)	(.25*)	(.35*)	(.58*)	(.40*)
Older adults	.75*	.76*	.78*	.66*	.87*	.83*	.68*	.77*	.78*	.69*	.74*	.73*	.39*[1]	.66*	.63*[1]	.45*
	(.75*)	(.76*)	(.77*)	(.67*)	(.87*)	(.79*)	(.68*)	(.79*)	(.78*)	(.67*)	(.72*)	(.76*)	(.39*[1])	(.64*)	(.60*[1])	(.44*)

Notes. ICCs are reported in brackets.
[1]Outlier removed.
*$p < .05$.

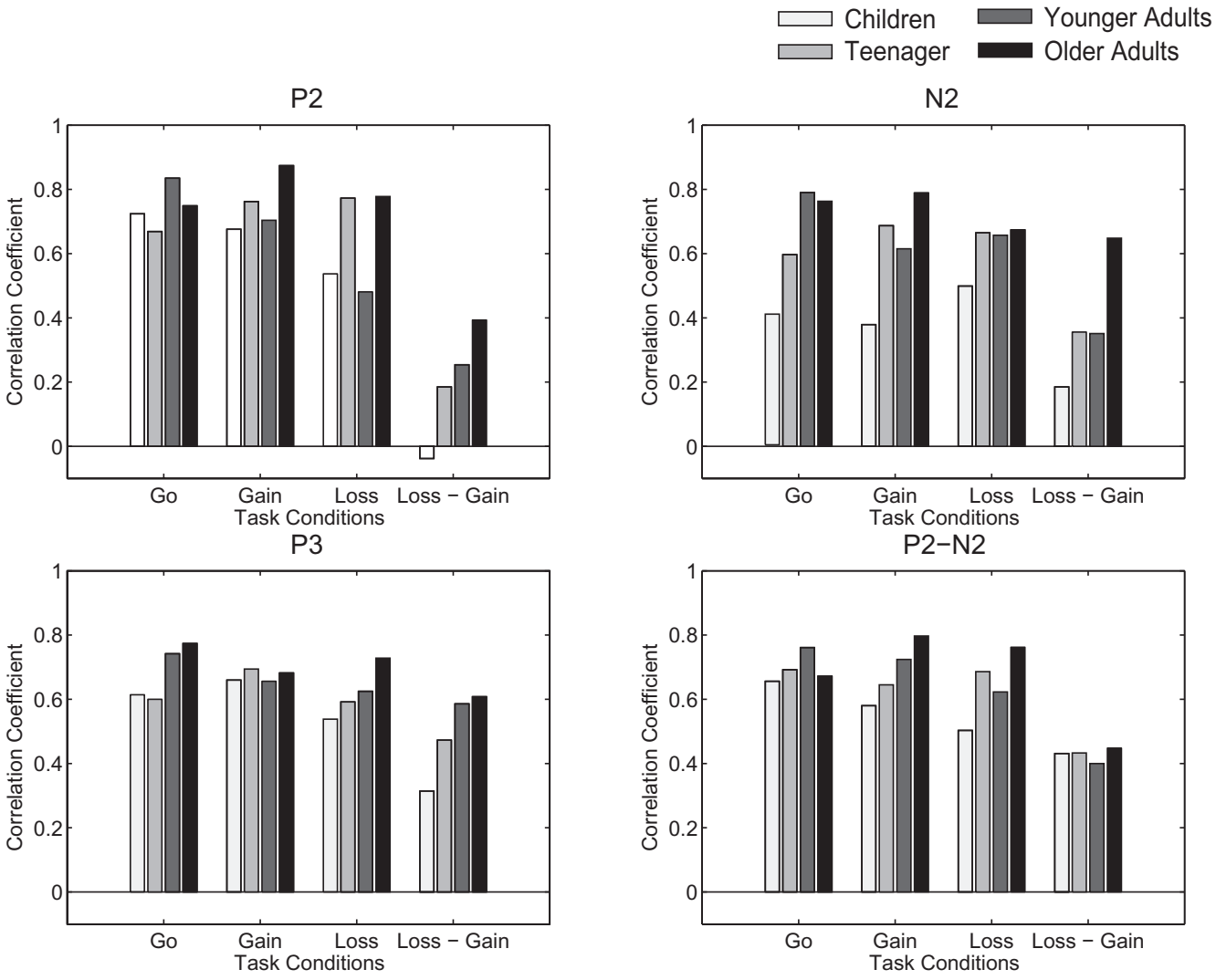

Figure 4. Intraclass test-retest correlations for the ERP amplitudes across the two testing sessions.

Table 3. *Multigroup Model Comparisons for Age Differences in Test-Retest Correlations of EEG Amplitudes*

	P2 Go		N2 Go		P3 Go		P2-N2 Go[1]		P2 Gain[1]		N2 Gain		P3 Gain		P2-N2 Gain[1]	
Model comparison	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI
Unconstrained—Equal correlations between age groups	.23	.002	2.94	.028	.88	.009	.32	.003	.84	.006	2.87	.032	.06	.001	.50	.005
	P2 Loss[1]		N2 Loss[1]		P3 Loss		P2-N2 Loss[1]		P2 Loss-gain		N2 Loss-gain		P3 Loss-gain[1]		P2-N2 Loss-gain	
Model comparison	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI	X^2diff	ΔNFI
Unconstrained—Equal correlations between age groups	2.5	.026	.51	.006	.54	.007	.88	.010	3.57	.344	3.25	.097	1.64	.033	.03	.001

[1]Outlier removed.

SNR Measures

SNR measures are shown in Figure 5. SNR measures were assessed for the four age groups for the Go-P3 amplitude. The SNR differed reliably between age groups, $F(3,94.6) = 8.27$, $p < .01$, $r_I = .46$. It was larger in younger adults and adolescents than in children and older adults (curvilinear contrast: $t = 4.47$, $p < .01$, $d = 0.88$). Furthermore, the expected increase in SNR with increasing number of trials, $F(19,135) = 92.76$, $p < .01$, $r_I = .96$, differed across age groups, $F(57,118) = 1.82$, $p < .01$, $r_I = .68$. As can be seen in Panel A of Figure 5, older adults started off with a SNR close to adolescents. However, the rise of SNR with increasing trial number was not as sharp as that of adolescents. With about 50 trials in the average, older adults' SNR curve was comparable to the consistently lower SNR of children. In younger adults, the SNR was consistently higher than in the other three age groups. To further investigate the reasons for the slower rise of SNR in children and older adults, we assessed the overall baseline noise level and signal power in the four age groups (cf. Panel B, Figure 5). Younger and older adults showed similar levels of baseline noise at all levels of aggregation, while children and adolescents showed higher noise levels than both groups of adults (contrast children and adolescents vs. younger and older adults: $t = 9.65$, $p < .01$, $d = 1.63$). Older adults are hence consistently less noisy than children. However, the signal is lower in older adults than in children (see inset Panel B, Figure 5). With increasing trial number, the higher noise level observed in children is compensated by the lower signal level in the case of the older adults, resulting in a slower rise of the SNR in older adults (cf. Panel A, Figure 5).

Discussion

The present study investigated the within-session reliabilities and test-retest stabilities of the P2, N2, and P3, as well as the P2-N2 peak-to-peak amplitude in the context of two performance monitoring tasks. Multigroup analyses were carried out to test for age differences in ERP test-retest stabilities among the four age groups. In addition, the SNR was assessed in each group to investigate whether age differences in stability or reliability might be due to age differences in signal or noise of the ERP measure. The main finding is that the amplitudes of stimulus-locked ERPs related to performance monitoring show comparable odd-even split reliability and test-retest stability across the lifespan. Satisfactory odd-

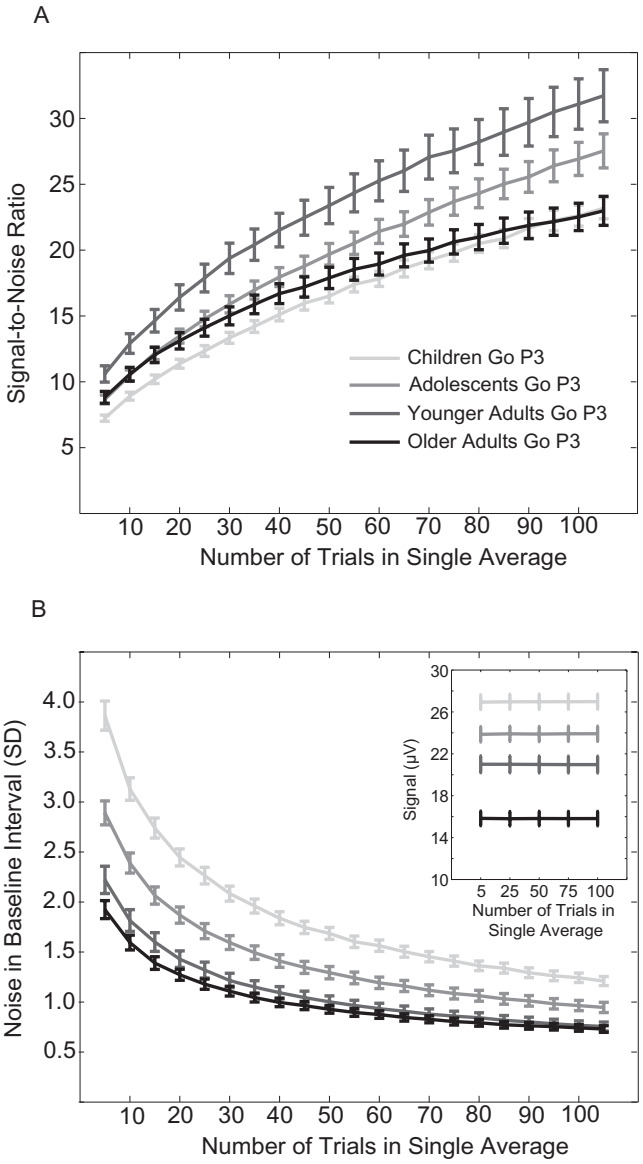

Figure 5. A: SNR ratio of the Go-P3 with increasing trial number in the single averages. B: Noise in 100-ms baseline interval (*SD*) and signal of the Go-P3 with increasing trial numbers (Inset). Error bars denote 1 *SE*.

even split reliabilities were observed in all four age groups with about 40 trials in the individual average. At the same time, SNRs differed across age groups, with children and older adults showing lower values than adolescents and younger adults. In children, the low SNR was driven by comparatively higher levels of noise, whereas in older adults, it was driven by comparatively low signal power.

Odd-Even Split and Split-Half Correlations as a Function of Trial Number

We observed satisfactory odd-even split reliabilities for the ERP amplitudes of the four ERPs investigated. In all age groups, odd-even split correlation coefficients reached an asymptote above .80. Hence, in all four age groups, satisfactorily low levels of measurement errors can be reached (for the Go-P3) when the number of trials in the individual average reaches or exceeds 40 trials. Further, split-half correlation coefficients reached a somewhat lower asymptote than odd-even split correlations in children, adolescents, and older adults, while this was less the case in younger adults. This suggests that interindividual differences in the change of the Go-P3 amplitude within a recording session were smaller in younger adults as compared to the other age groups. These age differences in the ratio of odd-even and split-half reliability suggest that the change of ERP amplitudes within a recording session differs more between individuals in the age groups of children, adolescents, and older adults as compared to younger adults. This finding is of importance for studies that investigate age differences in the development of ERPs during a recording session (e.g., when comparing learning effects). In these studies, an ERP amplitude decrease within a session, and also interindividual differences therein, should be normed for the different age groups using an ERP component that is assessed throughout the task but not supposed to be subject to learning effects. Also, this finding suggests that blocked condition comparisons should be avoided when comparing ERP measures across age groups.

Recent studies investigated the reliability of response-locked ERP components during performance monitoring and observed reliable estimates with as few as six trials in the average (Olvet & Hajcak, 2009a, 2009b; Pontifex et al., 2010), which stands in contrast to the 40 trials observed in this study. Given that stimulus-locked components appear about 200 ms after the critical event whereas response-locked ERPs appear about 100 ms later, these differences in reliability may be due to differences in the number of interfering processes. These, however, do not appear to differentially compromise the reliability across the lifespan.

Test-Retest Stabilities

The test-retest stabilities were overall lower than the odd-even split reliabilities, indicating variance related to individual changes across sessions in addition to the measurement error within a session. Also, corroborating the findings of prior studies, test-retest correlations assessed in Pearson product moment correlations and intraclass correlations differed only marginally (Segalowitz & Barnes, 1993; Thesen & Murphy, 2002). Taking into account changes at the group level between the two recording sessions, such as, for example, training effects, apparently did not contribute considerably to explaining differences in ERP amplitudes across the two task sessions. We would like to note that taking the average of the two test halves in an odd-even split measure or assessing the same task in two sessions both result in impoverished approximations of individual differences in intraindividual variability and change (Hertzog & Nesselroade, 2003). With respect to both measures, several split-up or repeated assessments would be preferable to obtain a more realistic approximation of variance related to intraindividual change within and across sessions. A first hint can be taken from a study that assessed the test-retest stability in an auditory oddball task in a series of eight sessions spaced across several months (Kinoshita, Inoue, Maeda, Nakamura, & Morita, 1996). Here, the ICCs ranged between .54 and .57 for the ERP amplitudes.

Age Differences in Test-Retest Stabilities

ERP amplitudes did not show systematic differences in test-retest stabilities between children, adolescents, younger adults, and older adults (cf. Figure 4). Together with the comparable odd-even split reliabilities across the age groups, this finding is encouraging, given the increasing number of developmental ERP studies in the domains of monitoring and executive control. It appears that ERP amplitudes with a minimum of 40 trials in the average can be assessed with comparable reliability in the different age groups, at least for the age ranges and components investigated here.

When comparing the size of the test-retest correlations across the age groups and ERPs, the amplitude of one ERP, the N2, seemed to be consistently less stable in children than in younger adults (see Figure 4). In light of the high and comparable odd-even split reliabilities for the N2 in all four age groups, this finding merits a comment. Prior evidence suggests that, in comparison to younger adults, children rely more on external than internal feedback to achieve behavioral control (Luna & Sweeney, 2004). Supporting this view, children exhibit larger cue-related ERPs or feedback-related ERPs, including the N2 in question, than adults in the context of both response conflict monitoring and outcome monitoring tasks (Eppinger, Mock, & Kray, 2009; Hämmerer, Li, Müller, & Lindenberger, 2010; Hämmerer et al., 2011; Jonkman, 2006). It might thus be the case that this stronger orientation to external stimuli in children as compared to the other age groups is reflected in less intraindividual stability across measurements in feedback-related ERPs.

SNR

As would be expected given their thinner skull (cf. Knott et al., 2004), we observed higher signal as well as noise levels in children and also in adolescents compared to the adult age groups for the Go-P3 (cf. Panel B, Figure 5). The ratio measure of signal levels relative to noise levels within each age group, however, is not affected by these functionally less relevant age differences in EEG signals. Independent of the number of trials in the average, the SNR was lower in children than in adolescents or younger adults, indicating that ERPs are inherently noisier during childhood. In contrast, the noise level in older adults was comparable to that of younger adults (cf. Panel B, Figure 5). This finding is somewhat unexpected, given that reaction time measures tend to become more variable with advancing adult age (cf. Li et al., 2004; MacDonald et al., 2006). However, things may be different for ERP data. The few data that exist on variability in older adults in ERP measures suggest no increase in latency jitter (Walhovd et al., 2008) or even reduced variability in older adults (Müller, Gruber, Klimesch, Lindenberger, 2009; Schmiedt-Fehr & Basar-Eroglu, 2011). One possible explanation might be that greater variability at the neural level with advancing adult age may result in flatter but

also more similar single-trial ERPs that ultimately vary less from each other. Older adults had a SNR comparable to that of children when the impact of noise was reduced by increasing trial numbers (cf. Panel A, Figure 5). With about 50 trials in the average, we observed comparable SNR levels in children and older adults, with children being compromised by high baseline noise levels and older adults by low signal power.

Comparison of Reliability, Stability, and SNR Measures: Recommendations for Age-Comparative ERP Studies

In line with the higher baseline noise level in children, odd-even split correlations were observed to be lower in children than in the other age groups. However, despite consistent age differences in SNR well beyond 40 trials in the average, comparable and reliable odd-even split correlations well above .80 can be observed in all four age groups with this number of trials in the average. The averaging procedure of ERPs can hence successfully reduce the impact of age differences in SNR, yielding a reliable estimate of the ERP amplitude in question. Thus, we recommend that future studies comparing age groups on stimulus-locked ERP measures should include a minimum of about 40 trials in the average.

The situation is different when the analyses are more sensitive to the SNR of ERPs, as is the case for dipole analyses or single-trial analyses. Here, it would be prudent to ascertain the SNR in each age group and strive for similar levels of SNR by increasing the number of trials in age groups with higher noise or lower signal levels. For example, in the present analyses of the SNR of the Go-P3, a SNR of about 20 would necessitate 35 trials per average in younger adults, 50 in adolescents, and 80 in children and older adults. However, we do not wish to evoke the impression that mechanisms contributing to age-based changes in baseline noise ought to be regarded as mere nuisance terms whose influences need to be minimized through age-differential aggregation. Instead, it has become increasingly clear that "noise" as captured in on-going EEG activity contributes to neural signaling (Deco, Jirsa, Robinson, Breakspear, & Friston, 2008) and evolves with age (McIntosh et al., 2010). Specifically, although EEG signals in the baseline interval do not reflect characteristics of ERP components and are commonly used as baseline in ERP research, it would be wrong to assume that these periods are completely task independent and are therefore reflecting pure EEG noise. However, it needs to be recognized that participants engage in preparatory cognitive processes during this interval, and that these processes may be reflected in aspects of the EEG signal that are more likely to manifest themselves in other types of EEG analysis (e.g., time frequency analyses). Hence, our recommendations are valid for certain applications on ERP research but do not necessarily generalize to other types of EEG analysis.

Finally, the lower split-half reliabilities in children, adolescents, and older adults indicate greater interindividual differences in the changes of ERP amplitudes during a recording session in these age groups compared to younger adults. This should be kept in mind when assessing effects that unfold in time during tasks such as learning-related processes or effects that are investigated across blocked conditions.

Conclusions and Outlook

The present study shows that ERPs related to performance monitoring exhibit high measurement reliabilities and moderate to high test-retest stabilities in children, adolescents, younger adults, and older adults. Multigroup analyses showed for the first time that amplitudes of ERPs related to performance monitoring did not differ between the four age groups. Due to the modest sample size, the present analyses may lack the statistical power[2] to detect small differences between age groups. Further research should overcome this limitation. Odd-even split correlations reached satisfactory levels in all age groups with about 40 trials in the average, even though SNR differed across age groups and were especially low in children and older adults. We conclude that the averaging of evoked responses successfully compensates for age differences in SNR. For analyses that are more sensitive to the SNR such as dipole analyses, researchers should consider adjusting the number of trials used for different age groups to yield comparable SNR levels. Also, when assessing changes in ERP amplitudes within a recording session, such as during learning, researchers should check for age-group differences in split-half reliabilities as a potential confound in data interpretation. Finally, various neurobiological factors are known or assumed to reflect in differences of EEG potentials across the lifespan, such as skull thickness, synaptic density, and myelination (e.g., Frodl et al., 2001, Picton & Taylor, 2007). These factors are most likely not spatially homogenous and might affect different EEG processes in different age groups. Further research, also in animal models, is needed to investigate how the multiplicity of age-graded changes in brain structure, chemistry, and function relate to differences in ERP reliability between age groups.

2. With the current sample size of about 180 participants, only medium and large effects ($d \geq .5$) can be detected with sufficient power ($1\text{-}\beta = .95$), thus it might be the case that smaller effects are present but have not been detected in the present study.

References

Bekker, E. M., Kenemans, J. L., & Verbaten, M. N. (2004). Electrophysiological correlates of attention, inhibition, sensitivity and bias in a continuous performance task. *Clinical Neurophysiology*, *115*, 2001–2013. doi: 10.1016/j.clinph.2004.04.008

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652. doi: 10.1037/0033-295X.108.3.624

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., . . . Reed, B. R. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General*, *130*, 746–763. doi: 10.1037/0096-3445.130.4.746

Brim, O. G., Jr., & Kagan, J. (1980). Constancy and change: A view of the issues. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 1–25). Cambridge, MA: Harvard University Press.

Bruin, K. J., & Wijers, A. A. (2002). Inhibition, response mode, and stimulus probability: A comparative event-related potential study. *Clinical Neurophysiology*, *113*, 1172–1182. doi: 10.1016/S1388-2457(02)00141-4

Campbell, K. B., Courchesne, E., Picton, T. W., & Squires, K. C. (1979). Evoked potential correlates of human information processing. *Biological Psychology*, *8*, 45–68. doi: 10.1016/0301-0511(79)90004-8

Cleary, T. A., & Linn, R. L. (1969) A note on the relative sizes of the standard errors of two reliability estimates. *Journal of Educational Measurement*, *6*, 25–27.

Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., & Friston, K. (2008). The dynamic brain: From spiking neurons to neural masses and cortical fields. *PLoS Computational Biology*, *4*, e1000092. doi: 10.1371/journal.pcbi.1000092

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Eppinger, B., Mock, B., & Kray, J. (2009). Insights into the development of reinforcement learning: Evidence from ERPs. *Psychophysiology*, *46*, 1–11. doi: 10.1111/j.1469-8986.2009.00838

Falkenstein, M., Hoormann, J., & Hohnsbein, J. (2002). Inhibition-related ERP components: Variation with modality, age, and time-on-task. *Journal of Psychophysiology*, *16*, 167–175. doi: 10.1027//0269-8803.16.3.167

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943. doi: 10.1126/science.1102941

Frodl, T., Meisenzahl, E. M., Muller, D., Leinsinger, G., Juckel, G., Hahn, K., . . . hegerl, U. (2001). The effect of the skull on event-related P300. *Clinical Neurophysiology*, *112*, 1773–1776.

Haig, A. R., Gordon, E., & Hook, S. (1997) To scale or not to scale: McCarthy and Wood revisited. *Electroencephalography and Clinical Neurophysiology*, *103*, 323–325. doi: 10.1016/S0013-4694(97)00009-6

Hämmerer, D., Li, S.-C., Müller, V., & Lindenberger, U. (2010). An electrophysiological study of response conflict processing across the lifespan: Assessing the roles of conflict monitoring, cue utilization, response anticipation, and response suppression. *Neuropsychologia*, *48*, 3305–3316. doi: 10.1016/j.neuropsychologia.2010.07.014

Hämmerer, D., Li, S.-C., Müller, V., & Lindenberger, U. (2011). Life span differences in electrophysiological correlates of monitoring gains and losses during probabilistic reinforcement learning. *Journal of Cognitive Neuroscience*, *23*, 579–592. doi: 10.1162/jocn.2010.21475

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*, 639–657. doi: 10.1037/0882-7974.18.4.639

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709. doi: 10.1037/0033-295X.109.4.679

Holroyd, C. B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: Electrophysiological responses to feedback stimuli. *Brain Research*, *1105*, 93–101. doi: 10.1016/j.brainres.2005.12.015

Jonkman, L. M. (2006). The development of preparation, conflict monitoring and inhibition from early childhood to young adulthood: A Go/NoGo ERP study. *Brain Research*, *1097*, 181–193. doi: 10.1016/j.brainres.2006.04.064

Jonkman, L. M., Lansbergen, M., & Stauder, J. E. A. (2003). Developmental differences in behavioral and event-related brain responses associated with response preparation and inhibition in a Go/NoGo task. *Psychophysiology*, *40*, 752–761. doi: 10.1111/1469-8986.00075

Joutsiniemi, S. L., Ilvonen, T., Sinkkonen, J., Huotilainen, M., Tervaniemi, M., Lehtokoski, A., . . . Näätänen, R. (1998). The mismatch negativity for duration decrement of auditory stimuli in healthy subjects. *Electroencephalography and Clinical Neurophysiology*, *108*, 154–159. doi: 10.1016/S0168-5597(97)00082-8

Kinoshita, S., Inoue, M., Maeda, H., Nakamura, J., & Morita, K. (1996). Long-term patterns of change in ERPs across repeated measurements. *Physiology & Behavior*, *60*, 1087–1092. doi: 10.1016/0031-9384(96)00130-8

Knott, D. P., Hazony, D., Karafa, M., & Koltai, P. J. (2004). High-frequency ultrasound in the measurement of pediatric craniofacial integrity. *Otolaryngology-Head and Neck Surgery*, *131*, 851–855. doi: 10.1016/j.otohns.2004.08.010

Labouvie, E. W. (1980). Identity versus equivalence of psychological measures and constructs. In L. W. Poon (Ed.), *Aging in the 1980s: Selected contemporary issues in the psychology of aging* (pp. 493–502). Washington, DC: American Psychological Association. doi: 10.1037/10050-036

Lamm, C., Zelazo, P. D., & Lewis, M. D. (2006). Neural correlates of cognitive control in childhood and adolescence: Disentangling the contributions of age and executive function. *Neuropsychologia*, *44*, 2139–2148. doi: 10.1016/j.neuropsychologia.2005.10.013

Li, S.-C., Hämmerer, D., Müller, V., Hommel, B., & Lindenberger, U. (2009). Lifespan development of stimulus-response conflict cost: Similarities and differences between maturation and senescence. *Psychological Research*, *73*, 777–785. doi: 10.1007/s00426-008-0190-2

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*, 155–163. doi: 10.1111/j.0956-7976.2004.01503003

Li, S.-C., von Oertzen, T., & Lindenberger, U. (2006). A neurocomputational model of stochastic resonance and aging. *Neurocomputing*, *69*, 1553–1560. doi: 10.1016/j.neucom.2005.06.015

Lövdén, M., Li, S.-C., Shing, Y. L., & Lindenberger, U. (2007). Within-person trial-to-trial variability precedes and predicts cognitive decline in old and very old age: Longitudinal data from the Berlin Aging Study. *Neuropsychologia*, *45*, 2827–2838. doi: 10.1016/j.neuropsychologia.2007.05.005

Luna, B., & Sweeney, J. A. (2004). The emergence of collaborative brain function: fMRI studies of the development of response inhibition. *Annals of the New York Academy of Sciences*, *1021*, 296–309. doi: 10.1196/annals.1308.035

MacDonald, S. W., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, *29*, 474–480. doi: 10.1016/j.tins.2006.06.011

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2008). Nobody is perfect: ERP effects prior to performance errors in musicians indicate fast monitoring processes. *PLoS ONE*, *4*, e5032. doi: 10.1371/journal.pone.0005032

McIntosh, A. R., Kovacevici, N., Lippe, S., Garrett, D., Grady, C., & Jirsa, V. (2010). The development of a noisy brain. *Archives Italiennes de Biologie*, *148*, 323–337.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798. doi: 10.1162/jocn.1997.9.6.788

Müller, V., Brehmer, Y., von Oertzen, T., Li, S. C., & Lindenberger, U. (2008). Electrophysiological correlates of selective attention: A lifespan comparison. *BMC Neuroscience*, *9*, 18.

Müller, V., Gruber, W., Klimesch, W., & Lindenberger, U. (2009) Lifespan differences in cortical dynamics of auditory perception. *Developmental Science*, *12*, 839–853. doi: 10.1111/j.1467-7687.2009.00834

Olvet, D. M., & Hajcak, G. (2009a). Reliability of error-related brain activity. *Brain Research*, *1284*, 89–99. doi: 10.1016/j.brainres.2009.05.079

Olvet, D. M., & 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

Papenberg, G., Bäckman, L., Chicherio, C., Nagel, I. E., Heekeren, H. R., Lindenberger, U., & Li, S.-C. (2011). Higher intraindividual variability is associated with more forgetting and dedifferentiated memory functions in old age. *Neuropsychologia*, *49*, 1879–1888. doi: 10.1016/j.neuropsychologia.2011.03.013

Picton, T. W., & Taylor, M. J. (2007). Electrophysiological evaluation of human brain development. *Developmental Neuropsychology*, *31*, 249–278.

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C.-T., Themanson, J., R., & Hillman, C. H. (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

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., Jr., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting & Clinical Psychology*, *20*, 343–350.

Sandman, C. A., & Patterson, J. V. (2000). The auditory event-related potential is a stable and reliable measure in elderly subjects over a 3 year period. *Clinical Neurophysiology*, *111*, 1427–1437. doi: 10.1016/S1388-2457(00)00320-5

Schmiedt-Fehr, C., & Basar-Eroglu, C. (2011) Event-related delta and theta brain oscillations reflect age-related changes in both a general and a specific neuronal inhibitory mechanism. *Clinical Neurophysiology*, *122*, 1156–1167. doi: 10.1016/j.clinph.2010.10.045

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, *30*, 451–459. doi: 10.1111/j.1469-8986.1993.tb02068

Segalowitz, S. J., Santesso, D. L., Murphy, T. I., Homan, D., Chantziantoniou, D. K., & Kan, S. (2010). Retest reliability of medial frontal negativities during performance monitoring. *Psychophysiology*, *47*, 260–270. doi: 10.1111/j.1469-8986.2009.00942

Thesen, T., & Murphy, C. (2002). Reliability analysis of event-related brain potentials to olfactory stimuli. *Psychophysiology*, *39*, 733–738. doi: 10.1111/1469-8986.3960733

Uwer, R., & von Suchodoletz, W. (2000). Stability of mismatch negativities in children. *Clinical Neurophysiology*, *111*, 45–52. doi: 10.1016/S1388-2457(99)00204-7

Walhovd, K. B., & Fjell, A. M. (2002). One-year test-retest reliability of auditory ERPs in young and old adults. *International Journal of Psychophysiology*, *46*, 29–40. doi: 10.1016/S0167-8760(02)00039-9

Walhovd, K. B., Rosquist, H., & Fjell, A. (2008). P300 amplitude age reductions are not caused by latency jitter. *Psychophysiology*, *45*, 545–553. doi: 10.1111/j.1469-8986.2008.00661

Winterer, G., & Weinberger, D. R. (2004). Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in Neurosciences*, *27*, 683–690. doi: 10.1016/j.tins.2004.08.002

(Received March 14, 2012; Accepted August 16, 2012)