Psychophysiology, •• (2012), ••–••. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2011.01349.x

BRIEF REPORT

Retest reliability of event-related potentials: Evidence from a variety of paradigms

SARAH M. CASSIDY, IAN H. ROBERTSON, AND REDMOND G. O'CONNELL
School of Psychology and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

Abstract

Event-related potentials (ERPs) offer unparalleled temporal sensitivity in tracing the distinct electrocortical processing stages enabling cognition and are frequently utilized in clinical and experimental investigations, yet few studies have investigated their retest reliability. We administered a battery of typical ERP paradigms to elicit a diverse range of components linked to distinct perceptual and cognitive processes (P1, N1, N170, P3a, P3b, error-related negativity, error positivity, P400). Twenty-five participants completed the battery on two occasions, 1 month apart. Analysis of component amplitudes indicated moderate-to-strong split-half and strong test-retest reliability. Peak latency reliability varied substantially across components and ranged from weak to strong. We confirm that a range of prominent ERPs provide highly stable neurophysiological indices of human cognitive function.

Descriptors: Cognition, Normal volunteers, EEG/ERP

Event-related potentials (ERPs) are a noninvasive method of measuring the distinct electrocortical processing stages enabling cognition and have become a popular tool in neuroscience, and increasingly in clinical and pharmacological investigations. The validity of ERPs as endophenotypes or as biomarkers is partly dependent on their stability over time. Although ERPs are modulated state factors including circadian rhythm (Munte, Heinze, Kunkel, & Scholz, 1987), sleep deprivation (Boonstra, Stins, Daffertshofer, & Beek, 2007; Murphy, Richard, Masaki, & Segalowitz, 2006), and mood (e.g., Cavanagh & Geisler, 2006; Olvert & Hajcak, in press), they should exhibit substantial consistency over time if they index stable neurophysiological traits.

To date, retest evaluations have focused on a limited range of ERP components, primarily the P300 (e.g., Walhovd & Fjell, 2002), mismatch negativity (MMN, e.g., Hall et al., 2006), and error-related negativity (ERN, e.g., Segalowitz et al., 2010) and have reported moderate-to-strong reliability estimates over periods ranging from weeks to years. Our study expands on this work by calculating retest reliability estimates for a diverse range of commonly studied ERP components collected from a single sample of participants. This approach facilitated comparison of the relative stability of different categories of components (e.g., early vs. late, perceptual vs. cognitive) and component measures (e.g., mean amplitude vs. area-under-the-curve, absolute peak vs. difference waveform).

We employed paradigms that are frequently utilized in the literature to elicit the components of interest. Task selection was designed to ensure a broad sample of ERPs indexing a range of different perceptual and cognitive processes, including visual-evoked components P1 and N1 (early visual selection, Luck, Woodman, & Vogel, 2000), the N170 (face processing, Ming-Fang, Ye, & Qing-Lin, 2010), the P3a and P3b (attention resource allocation, Polich & Criado, 2006), the ERN and Pe (error processing, Falkenstein, Hoormann, Christ, & Hohnsbein, 2000), and the P400 (memory encoding, Dien, Michelson, & Franklin, 2010). The current study examines the test-retest and split-half reliability of these commonly recorded ERP components across a 1-month period.

Method

Participants

Twenty-five right-handed participants (15 females, 19–35 years old; mean age, 24.6; *SD*, 4.5) with normal or corrected-to-normal vision and no history of head injury, neurological or psychiatric disorder. Participants gave written informed consent prior to the study, which was approved by the Trinity College Dublin School of Psychology Ethics Committee, and each received €40 for participating.

Two participants were excluded from the Oddball analysis and one from the Sternberg analysis due to incomplete data collection arising from a technical fault. One participant was

This research was supported by an Irish Research Council for Science, Engineering and Technology (IRCSET) Empower Fellowship to R.G.O'C. The authors also acknowledge funding support via the HEA PRTLI Cycle 3 program of the EU Structural Funds and the Irish Government's National Development Plan 2002–2006. This work was conducted in association with NIEL (Neuroenhancement for Inequalities in Elder Lives), a research program funded by Atlantic Philanthropies.

Address correspondence to: Dr. Redmond O'Connell, Trinity College Institute of Neuroscience, Lloyd Building, Trinity College Dublin, Dublin 2, Ireland. E-mail: reoconne@tcd.ie

removed from the Flanker analysis due to insufficient trial numbers after artifact rejection (<6 errors). Two participants were removed from the Faces analysis: one due to insufficient trial numbers and one due to excessive artifacts. In addition, any participants with extreme outlier values on a given component were excluded from the reliability estimates for that component only. Final participant numbers for each reliability estimate are provided in Table 1.

Procedure

Participants performed four cognitive tests on two separate occasions separated by 4 weeks. Each session lasted approximately 90 min including frequent rest breaks. Each of the four tests is explained below. In all cases, participants fixated on a central cross, and speed and accuracy were equally emphasized. Participants were tested while seated in an armchair ~65 cm from the computer screen in a dimly lit, electrically shielded room. Both testing sessions took place at the same time of day. All participants completed their second session 28 ± 2 days after the first session.

Oddball task (P1, N1, P3a, P3b). Standard stimuli consisted of 3.5-cm diameter purple circles, which appeared on 75% of trials. Target stimuli were 4-cm diameter purple circles, which appeared on 12.5% of trials. Distractor stimuli were black and white checkerboards, which appeared on 12.5% of trials. Every 2,075 ms, a stimulus appeared on the screen for 75 ms. Participants indicated Target stimuli by pressing the left mouse button. There were 320 trials in total.

Flanker task (ERN, Pe). Five-letter arrays, in which a central target letter was flanked on each side by either congruent (SSSSS or HHHHH) or incongruent (HHSHH, SSHSS) letters, were serially presented at fixation. Participants made a left click if the central stimulus was H and a right click if the central stimulus was S. There were 480 trials in total; 80 congruent array trials and160 incongruent array trials for both of the letters. Each array was on screen for 200 ms with an interstimulus interval of 1,250 ms.

Sternberg task (P1, N1, P400). Participants were presented with 100 different memory sets, consisting of five digits serially presented at fixation (stimulus duration 500 ms, interstimulus interval 1,000 ms), to be retained in working memory during each trial. A single probe digit was presented 3 s after the last set item, and participants pressed the left mouse button once if the probe had been absent (negative probe) or twice if it was present (positive probe) from the preceding set of numbers. The probability that the probe was a member of the preceding memory set was 0.5.

Faces task (N170). Stimuli consisted of 100 faces, of which 50 were male (25 neutral expression, 25 happy) and 50 were female (25 neutral expression, 25 happy). Participants pressed either the left or right mouse button (indicated by a response cue presented after each face) according to face gender. The response cue consisted of the letter M and the letter F randomly presented either side of fixation to prevent motor preparation. Each trial consisted of a 500-ms blank screen, followed by the face stimulus (70 ms) followed by another 500-ms blank screen, followed by the response cue, which stayed on screen until a response was registered. Identical face stimuli were presented at time 1 and time 2.

Assessing state factors. Participants completed a brief questionnaire assessing the number of hours of sleep the night before each testing session and current levels of stress and energy (single question: "How much energy would you say you have right now?" responding on a 5-point Likert scale). The Hospital Anxiety and Depression Scale (HADS; Zigmond & Sims, 1983) was also administered. Participants were asked to abstain from caffeine and nicotine 2 h before testing.

EEG Acquisition

Continuous electroencephalogram (EEG) was acquired through the ActiveTwo BioSemi electrode system from 64 scalp electrodes, digitized at 512 Hz. Vertical eye movements were recorded with two vertical electrooculogram electrodes placed above and below the left eye, while electrodes at the outer canthus of each eye recorded horizontal movements. EEG preprocessing and analysis were conducted using BESA 5.2. Data were average referenced offline and filtered from 0.5–40 Hz. Blinks and eye movements were corrected using an algorithm developed by Berg and Scherg (1994). Individual epochs were subjected to an additional artifact criterion of ±90 mV.

Stimulus-locked ERPs were extracted from the Oddball, Sternberg, and Faces tasks and segmented into epochs of 100 ms before to 800 ms after stimulus onset, and baseline-corrected using the prestimulus interval. Scalp locations and measurement windows for each ERP component were based on their spatial extent and latency after inspection of grand-average waveforms (collapsed across the two sessions). Later, ERP components P3a/P3b, Pe, and P400, which do not have well-defined peaks, were measured using a variety of common scoring methods (peak amplitude, mean amplitude, area-under-the-curve). To limit the number of statistical comparisons, reliability estimates were calculated for the electrode location at which component amplitude was greatest (see Table 1). The P1 (70–110 ms), N1 (130–180 ms), P3a (300–450 ms) were elicited by distractor stimuli on the Oddball task, the P3b (300–500 ms) was elicited by target stimuli on the Oddball task, memory set stimuli on the Sternberg task elicited the P1 (70–130 ms) and N1 (130–210 ms), and positive probe stimuli were used to elicit the P400 (350–450 ms). A comparison of happy versus neutral faces revealed no significant effects on N170 amplitude or latency at either session (all $t < 1$). Nevertheless, separate N170 (120–190 ms) measurements were obtained for happy versus neutral face stimuli.

The ERN and Pe, elicited by erroneous responses during the Flanker task, are response-locked components and were calculated by averaging 200 ms before to 600 ms after erroneous responses and baseline-corrected relative to the preresponse interval. The ERN (0–100 ms) was measured in the form of peak amplitude (maximum negative voltage 0–100 ms) and peak-to-peak (subtracting negative peak from maximum positive voltage –100 ms to 0). Pe amplitude was measured in the interval of 140–280 ms postresponse.

ERN/Pe and P3a/P3b component measures were also extracted from difference waveforms designed to highlight task-specific effects. For the ERN/Pe, a difference waveform was calculated by subtracting the average ERP elicited by standard go trials from the average ERP elicited by errors for each participant. For the P3a, a difference waveform was generated by subtracting standard trials from distractor trials and for the P3b, by subtracting standard trials from target trials. The same time windows detailed above were used to measure difference waveform components.

Table 1. *Summary of Behavioral and Electrophysiological Results for Sessions 1 and 2*

Behavioral analysis					ERP test-retest analysis								ERP split-half (Spearman Brown corrected)			
		mean T1 (SD)	*mean T2 (SD)*	*T*		Site	*n*	mean *T1 (SD)*	mean *T2 (SD)*	*T*	*r*	ICC	Site	*n*	*r*	ICC
Oddball	%acc	77.4 (19.8)	80.5 (15.5)	–1.1	P1											
	RT	499 (98)	480 (83)	1.4	Peak Amplitude	P08	21	5.9 (3.5)	5.6 (3.3)	0.5	0.75***	0.76***	P08	23	0.69**	0.61**
					Peak Latency	P08	21	92 (8.4)	91 (9.0)	0.4	0.38	0.39*	P08	23	0.74***	0.73***
					N1											
					Peak Amplitude	P08	21	–7.3 (4.4)	–8.3 (4.5)	2.1*	0.89***	0.87***	O2	22	0.88***	0.87***
					Peak Latency	P08	21	150 (12)	149 (10)	0.52*	0.52*	0.53*	O2	22	0.86***	0.87***
					P3a											
					Peak Amplitude	Cz	22	7.2 (4.0)	6.9 (3.8)	0.8	0.77***	0.77***	Cz	23	0.93***	0.90***
					Mean Amplitude	Cz	22	3.9 (2.8)	3.8 (3.1)	0.5	0.78***	0.78***	Cz	23	0.90***	0.86***
					Area Under Curve	Cz	22	869 (500)	846 (541)	0.3	0.77***	0.78***	Cz	23	0.89***	0.86***
					Peak Latency	Cz	22	364 (37)	390 (51)	–2.6*	0.46*	0.38	Cz	23	0.38	0.38
					Difference Peak	Cz	22	3.6 (2.9)	3.1 (2.9)	–0.5	0.64**	0.64**	Cz	23	0.66***	0.63***
					Difference Peak Latency	Cz	22	358 (31)	362 (35)	–1	0.89***	0.88***	Cz	23	0.34	0.33
					Difference Mean Amp	Cz	22	3.8 (2.4)	3.4 (2.6)	–1.1	0.83***	0.82***	Cz	23	0.70***	0.66***
					Difference Area	Cz	22	579 (342)	343 (77)	–0.8	0.82***	0.82***	Cz	23	0.73***	0.72**
					P3b											
					Peak Amplitude	CPz	22	5.4 (2.4)	6.4 (3.01)	–2.8*	0.84***	0.77***	CPz	23	0.73***	0.62**
					Mean Amplitude	CPz	22	3.9 (2.2)	4.6 (2.6)	–2.2*	0.83***	0.8***	CPz	23	0.68**	0.53*
					Area Under Curve	CPz	22	810 (406)	929 (504)	–2.3*	0.88***	0.83***	CPz	23	0.63*	0.49*
					Peak Latency	CPz	22	427 (66)	421 (47)	0.4	0.42*	0.41*	CPz	23	0.01	0.01
					Difference Peak	CPz	22	2.4 (2.8)	2.8 (3.1)	0.6	0.55**	0.52**	CPz	23	0.63***	0.61***
					Difference Peak Latency	CPz	22	426 (53)	421 (43)	0.4	0.18	0.18	CPz	23	0.03	0.03
					Difference Mean Amp	CPz	22	1.5 (2.4)	1.6 (2.7)	0.3	0.73***	0.73***	CPz	23	0.71***	0.70***
					Difference Area	CPz	22	497 (331)	582 (286)	1.5	0.61**	0.59**	CPz	23	0.74***	0.74**
Flanker	Error trials	16.1 (11.5)	16.8 (9.5)	–0.6	ERN											
	RT error trials	457 (86)	469 (107)	0.6	Peak Amplitude	FCz	22	–7.6 (3.6)	–7.7 (2.9)	–0.8	0.75***	0.74***	FCz	20	0.64*	0.64*
					Peak-to-peak	FCz	22	10.3 (4.9)	10.9 (4.4)	0.4	0.76***	0.76***	FCz	20	0.51*	0.52**
					Peak Latency	FCz	22	49 (15)	52 (17)	–0.8	0.43*	0.45*	FCz	20	0.39	0.39
					Difference Peak	FCz	22	–7.2 (4.3)	–7.0 (4.3)	–0.4	0.82***	0.87***	FCz	20	0.72***	0.69***
					Difference Peak Latency	FCz	22	59 (19)	54 (20)	1	0.31	0.3	FCz	20	0.15	0.15
					Difference Peak-to-peak	FCz	22	–2.2 (1.7)	–2.3 (1.3)	–0.4	0.57**	0.56**	FCz	20	0.44	0.38
					Pe											
					Peak Amplitude	FCz	22	6.3 (4.8)	7.1 (3.4)	–1.4	0.74***	0.71***	FCz	20	0.88***	0.86***
					Mean Amplitude	FCz	22	3.4 (3.2)	4.0 (2.6)	–1.1	0.63**	0.62**	FCz	20	0.76***	0.72***
					Area Under Curve	FCz	22	507 (395)	576 (269)	–0.99	0.58**	0.54**	FCz	20	–0.08	–0.08
					Peak Latency	FCz	22	202 (34)	191 (40)	1.3	0.44*	0.43*	FCz	20	0.52*	0.49*
					Difference Peak	FCz	22	8.5 (5.6)	8.6 (4.4)	0.2	0.87***	0.85***	FCz	20	0.89***	0.89***
					Difference Peak Latency	FCz	22	219 (31)	207 (38)	–1.4	0.4	0.39	FCz	20	0.3	0.3
					Difference mean Amp	FCz	22	4.9 (3.9)	5.5 (3.5)	1	0.75***	0.74***	FCz	20	0.85***	0.85***
					Difference Area	FCz	22	781.2 (479.2)	800.9 (456,4)	0.3	0.78***	0.78***	FCz	20	0.84***	0.84***
Sternberg	% acc	97.8 (1.9)	96.6 (5.3)	1.1	P1											
	RT	873 (227)	814 (191)	1.8	Peak Amplitude	P08	24	2.6 (1.6)	2.4 (1.6)	0.8	0.76***	0.76***	P08	24	0.76***	0.54**
					Peak Latency	P08	24	98 (14)	97 (12)	0.8	0.78***	0.77***	P08	24	0.65**	0.66**
					N1											
					Peak Amplitude	P07	24	–5.6 (3.5)	–5.7 (3.7)	0.5	0.91***	0.91***	P07	23	0.89***	0.77***
					Peak Latency	P07	24	156 (14)	158 (17)	–1.1	0.89***	0.87***	P07	23	0.93***	0.81***
					P400											
					Peak Amplitude	Cz	24	3.1 (3.1)	2.8 (2.6)	0.9	0.84***	0.85***	Cz	24	0.87***	0.87***
					Mean Amplitude	Cz	24	1.9 (3.2)	1.7 (2.6)	0.8	0.86***	0.85***	Cz	24	0.88***	0.88***
					Area Under Curve	Cz	24	287 (234)	243 (196)	1.7	0.83***	0.80***	Cz	24	0.73***	0.68***
					Peak Latency	Cz	24	400 (37)	392 (38)	0.8	0.19	0.19	Cz	24	0.05	0.06
Faces	Mean acc	94.9 (6.7)	97.4 (2.0)	–1.2	N170											
	RT	528 (78)	521 (78)	1.7	Happy Faces											
					Peak Amplitude	P10	23	–7.0 (4.0)	–6.8 (3.8)	–0.5	0.82***	0.83***	P10	21	0.81***	0.81***
					Peak Latency	P10	23	143 (18)	142 (13)	0.3	0.78***	0.75***	P10	21	0.83***	0.77***
					Neutral Faces											
					Peak Amplitude	P10	23	–6.7 (3.2)	–6.7 (3.9)	–0.2	0.85***	0.84***	P10	21	0.80***	0.80***
					Peak Latency	P10	23	140 (16)	144 (18)	–1.9	0.79***	0.77***	P10	21	0.90***	0.90***

Notes. Test-retest and split-half reliability estimates (Pearson's correlation coefficient, *r*, and intraclass correlation coefficient, ICC) are provided for each ERP component with separate values for each peak measurement method used.
*$p < .05$. **$p < .01$. ***$p < .001$.

The proportion of participants exhibiting voltage values in the expected direction was calculated to establish the prevalence of P1, N1, N170, and P400 components. For the ERN and Pe, a component was deemed to be present if the participant had a larger peak voltage on error versus correct trials. For the P3a and P3b, target and distractor trials were compared to standard trials in order to assess the presence of a component in each participant.

Statistical Analysis

Test-retest reliability indices for ERP measures were obtained both in terms of intersubject stability (Pearson's correlation coefficient, *r*) and score agreement (intraclass correlation coefficients, ICC). ICCs reflect the consistency of a measure taking into account variance related to the time of testing (Shrout & Fleiss, 1979), whereas the Pearson's correlation coefficient reflects intersubject stability according to subjects' ranking. Split-half reliability was also performed by comparing the first half of trials in session 1 to the second half in session 1. Because split-half reliability metrics were based on half of the trials, these measures were corrected using the Spearman–Brown prophecy formula (Helmstadter, 1964). *T* tests were performed to compare behavioral data and ERP components between session 1 and session 2. In addition, to explore within-subjects relationships between components, Pearson correlations were calculated between related ERP component pairs measured at time 1. In addition to the P1/N1 measured on the Oddball and Sternberg tasks, we explored the correlation between the Pe, P3a, and P3b based on previous proposals that these components reflect the same psychophysiological process (Overbeek, Nieuwenhuis, & Ridderinkhof, 2005). Before calculating reliability estimates, all variables were checked for extreme outlier values, and any participants with values that were greater than 3 standard deviations above the mean were removed from the analysis in question. In all cases, the removal of extreme outliers resulted in lower correlation coefficients.

Results

There were no significant differences in the amount of sleep, $t(24) = 0.2, p > .05$; energy, $t(24) = -0.5, p > .05$; stress, $t(24) = 1.5$, $p > .05$; anxiety, $t(24) = -1.0$, $p > .05$, or depression $t(24) = 0.2$, $p > .05$, at time 1 versus time 2.

All ERP components were evident in the overwhelming majority of participants (average: 97%, range: 83% (P400)—100% (P1/N1)) with no significant differences across sessions, $t(9) = -0.48$, $p > .05$.

Complete retest reliability results are presented in Table 1 and Figure 1. There were no significant differences in behavioral performance across sessions. We observed moderate-to-strong split-half and strong test-retest reliability for all component amplitudes with some variation across measurement techniques. Reliability estimates for latency measures were in the weak-to-moderate range for the longer latency components (P3a/P3b, Pe, P400) but remained strong for the N170 and for P1 and N1 components elicited during the Sternberg task. Only two components (Oddball N1 and P3b) showed a change in magnitude from session 1 to session 2.

The correlation between P1 components measured on the Oddball and Sternberg tasks did not reach significance, but a strong relationship was observed for the N1 ($r = 0.7$, $p < .001$). In addition, the Pe was found to correlate with the P3b ($r = 0.43$, $p < .05$) and not the P3a ($r = 0.3$, $p > .05$).

Discussion

Our data consistently indicate strong retest reliability at follow-up for amplitude measurements of P1, N1, N170, P3a, P3b, ERN, Pe, and P400, and this stability was evidenced across all statistical comparisons (retest/split-half, Pearson's *r*/ICC) with some variation depending on the scoring method that was used. Reliability estimates also remained high when selected ERPs (ERN, Pe, P3a, P3b) were measured from difference waveforms that isolated the task-specific effects. In keeping with most previous studies, we found that peak latency also exhibited significant stability across sessions, but reliability indices were in the small-to-moderate range in most cases with the exceptions of the P1/N1 and the N170. Although the impact of component scoring methods was limited, mean amplitude and area-under-the-curve were the most stable measurement methods for the P3a, P3b, and P400 while peak amplitude was most stable for the shorter latency components (P1, N1, N170, ERN). Strong intertask ERP correlations were also observed for the N1 (Oddball and Sternberg) and between the Pe and P3b. This last finding supports the proposal that the Pe and P3b index the same psychophysiological process (Overbeek, Nieuwenhuis, & Ridderinkhof, 2005).

The existing ERP reliability literature lacks consistency in terms of the components that are investigated and the retest intervals that are used. We are not aware of any previous studies that have evaluated the reliability of the visual P1/N1, N170, P3a, or P400, which precludes any comparison for these components, but several previous studies of the ERN, Pe, and P3b have yielded comparable reliability estimates.

It is noteworthy that our results suggest comparable reliability for early perceptual components like the P1, N1, and N170 as for the higher level cognitive components. P1 and N1 reliability was substantially reduced on the Oddball task compared to the Sternberg task, but this is likely due to the difference in trial numbers available on each task and the consequent impact on signal/noise ratios (maximum 40 trials on Oddball vs. 500 on Sternberg).

In considering the reliability of ERP components, it is important to draw a distinction between the absolute amplitude and latency of a component elicited by a given stimulus and task-specific ERP effects that are best isolated by computing difference waveforms. In the present study, we found that ERP reliabilities were maintained even after isolating task-specific effects (ERN, Pe, P3a, P3b). A limitation of this and other similar studies is that such contrasts were not possible for the N170 and P400 due to the task designs, and future research is required to address this discrepancy (e.g., N170 face vs. non-face stimulus, P400 high vs. low memory load).

The present study confirms that the selected range of ERP components, associated with a diverse range of perceptual and cognitive functions, represent a highly stable neurophysiological index of cognitive function in neurologically healthy adults up to 1 month after initial acquisition. ERPs derived from these tasks therefore represent a suitable tool for investigations of personality traits, endophenotype models, and clinical trials.

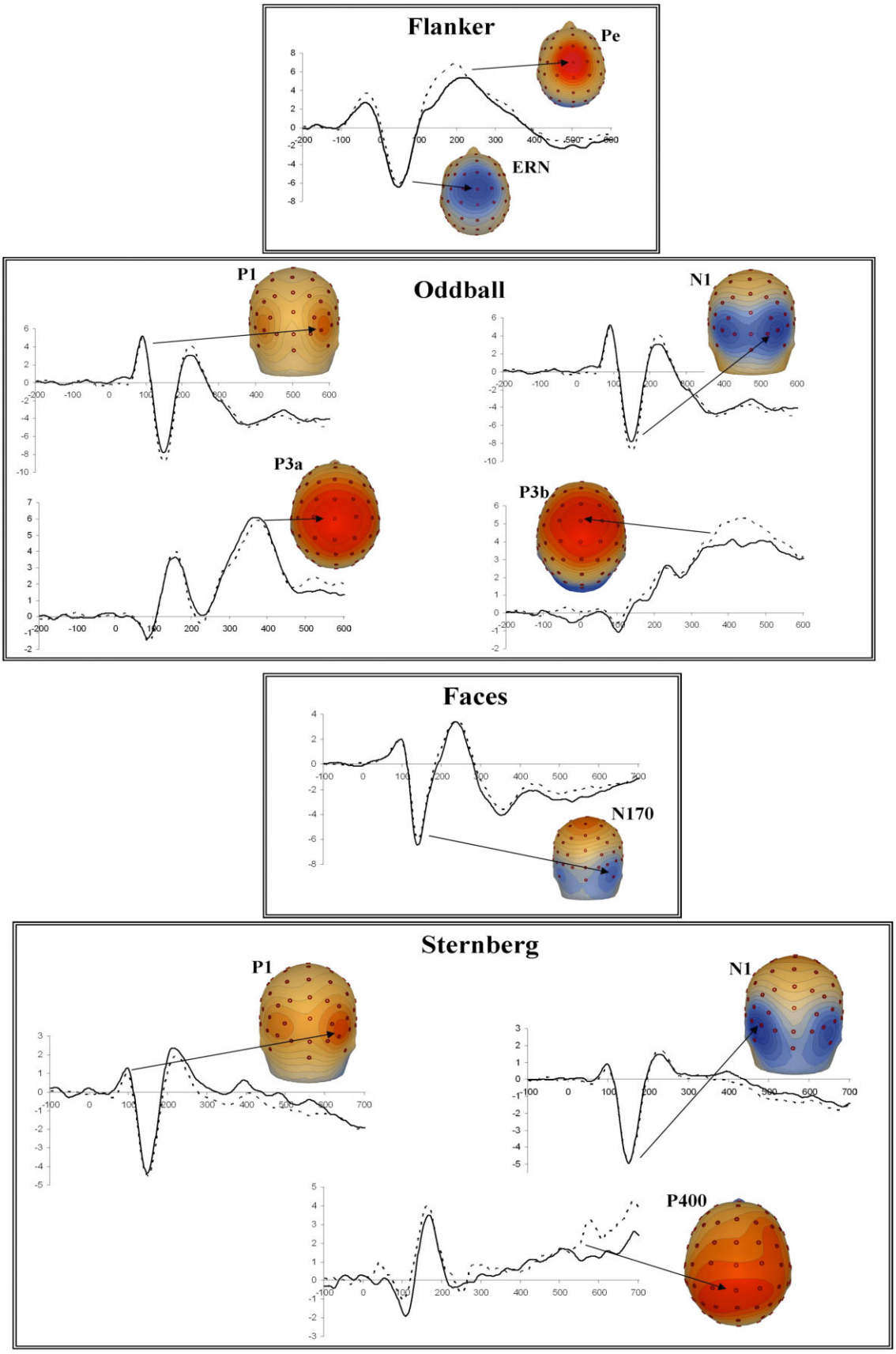

Figure 1. Grand-average ERP waveforms and associated component topographies at initial testing (dark line) and at 1 month follow-up (dashed line).

References

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography & Clinical Neurophysiology 90*, 229–241.

Boonstra, T. W., Stins, J. F., Daffertshofer, A., & Beek, P. J. (2007). Effects of sleep deprivation on neural functioning: An integrative review. *Cellular and Molecular Life Sciences, 64*, 934–946. doi: 10.1007/s00018-007-6457-8

Cavanagh, J., & Geisler, M. W. (2006). Mood effects on the ERP processing of emotional intensity in faces: A P3 investigation with depressed students. *International Journal of Psychophysiology, 60*, 27–33. doi: 10.1016/j.ijpsycho.2005.04.005

Dien, J., Michelson, C. A., & Franklin, M. S. (2010). Separating the visual sentence N400 effect from the P400 sequential expectancy effect: Cognitive and neuroanatomical implications. *Brain Research, 1355*, 126–140.

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology, 51*, 87–107.

Hall, M. H., Schulze, K., Rijsdijk, F., Picchioni, M., Ettinger, U., Bramon, E., . . . Sham, P. (2006). Heritability and reliability of P300, P50 and duration mismatch negativity. *Behavior Genetics, 36*, 845–857. doi: 10.1007/s10519-006-9091-6

Helmstadter, G. C. (1964). *Principles of psychological measurement*. New York, NY: Appleton-Century Crofts.

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Science, 4*, 432–440. doi: S1364-6613(00)01545-X

Ming-Fang, L., Ye, Z., & Qing-Lin, Z. (2010). A review of the N170 component in face recognition. *Advances in Psychological Science*, *18*, 1942–1948.

Munte, T. F., Heinze, H. J., Kunkel, H., & Scholz, M. (1987). Human event-related potentials and circadian variations in arousal level. *Progress in Clinical and Biological Research*, *227B*, 429–437.

Murphy, T. I., Richard, M., Masaki, H., & Segalowitz, S. J. (2006). The effect of sleepiness on performance monitoring: I know what I am doing, but do I care? *Journal of Sleep Research*, *15*, 15–21. doi: 10.1111/j.1365-2869.2006.00503.x

Olvert, D. M., & Hajcak, G. (in press). The error-related negativity relates to sadness following mood induction among individuals with high neuroticism. *Social Cognitive and Affective Neuroscience*.

Overbeek, T. J. M., Nieuwenhuis, S., & Ridderinkhof, K. R. (2005). Dissociable components of error processing: On the functional significance of the Pe vis-à-vis the ERN/Ne. *Journal of Psychophsiology*, *19*, 319–329.

Polich, J., & Criado, J. R. (2006). Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology*, *60*, 172–185. doi: 10.1016/j.ijpsycho.2005.12.012

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

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: S0167876002000399

Zigmond, A. S., & Sims, A. C. (1983). The effect of the use of the international classification of diseases 9th revision: Upon hospital in-patient diagnosis. *British Journal of Psychiatry*, *142*, 409–413.

(Received June 8, 2011; Accepted December 3, 2011)