Article

Test-Retest Reliability of Event-Related Potentials Across Three Tasks

Simon Morand-Beaulieu[1,2,3], Marie-Ange Perrault[2,4], and Marc E. Lavoie[2,3,5]

[1]Child Study Center, Yale University School of Medicine, New Haven, CT, USA
[2]Laboratoire de psychophysiologie cognitive et sociale, Centre de recherche de l'Institut universitaire en santé mentale de Montréal, QC, Canada
[3]Département de neurosciences, Université de Montréal, QC, Canada
[4]Département de psychologie, Université de Montréal, QC, Canada
[5]Département de psychiatrie et d'addictologie, Université de Montréal, QC, Canada

Abstract: Event-related potentials (ERPs) constitute a useful and cost-effective method to assess the neural underpinnings of multiple cognitive processes. ERPs have been used to track changes in cognitive processes in longitudinal and clinical studies. However, few studies have assessed their test-retest reliability (i.e., their consistency across time). Therefore, in the current study, we aimed to assess the test-retest reliability of ERPs (P100, N100, P200, N200, P3b, lateralized readiness potentials) across three tasks. In two assessments separated by approximately 4 months, ERPs were recorded in 26 healthy participants, during two oddball tasks (motor and counting) and a stimulus-response compatibility paradigm. Pearson's correlations and intraclass correlations were used to assess the test-retest reliability of ERPs. Correlations between ERPs elicited by the three tasks were assessed with Pearson's correflations. Our analyses revealed moderate to very strong test-retest reliability for most ERP components across the three tasks. Test-retest reliability did not differ between the motor and counting oddball tasks. Most ERPs were also correlated across paradigms. Therefore, these results confirm that ERPs have the potential to be reliable markers to serve as robust assessment tools in longitudinal or clinical studies.

Keywords: event-related potentials, cognition, oddball, stimulus-response compatibility, test-retest reliability

Event-related potentials (ERPs) are obtained through the time-locked averaging of a continuous electroencephalogram (EEG) recording. From this averaged ERP signal, one can derive specific components, which represent the electrical brain response to a given cognitive occurrence (Luck, 2005). ERPs allow a very precise temporal tracking of the electrocortical brain activity, which in turn allows high sensitivity to the rapid information processing stream. Given their use to track cognitive processes in longitudinal (Fruehwirt et al., 2018; Wachinger et al., 2018) or clinical studies (Houston & Schlienz, 2018; Morand-Beaulieu et al., 2018), the test-retest reliability of ERPs must be established. Test-retest reliability refers to the consistency of a measure or an instrument over time. Thus, if we were to assess an informant with an instrument that has good test-retest reliability at different time points, that instrument would yield similar scores at all time points. In the context of ERP research, an ERP component with good test-retest reliability would have constant amplitude and latency over time.

The oddball task is one of the most frequently used tasks in ERP research (Kutas et al., 2012). This task involves the presentation of a stream of standard stimuli, which is disrupted at times by the presentation of a deviant stimulus. This procedure is known to elicit the P300 (or P3b), which is one of the most studied ERP components. The P3b has a maximal amplitude over parietal electrodes and has generators distributed across the cortex, including notably the parietal, temporal, and posterior cingulate cortices (Bledowski et al., 2004; Morgan et al., 2016; Polich, 2007). Some versions of the oddball task require a button press when deviant stimuli are presented while others require a silent count of deviant stimuli. Investigating ERPs in both versions of the task allows us to isolate the effect of button-pressing on components of interest (e.g., the P3b). However, the literature regarding the effects of button-pressing on ERP components in oddball tasks remains inconsistent (Ford et al., 2000; Kayser et al., 2010; Kok, 1988; Salisbury et al., 2001; Wronka et al., 2008). Such comparison of both oddball task variants has been used in

clinical studies. For instance, following cognitive-behavioral therapy, individuals with Tourette syndrome or body-focused repetitive behaviors were found to have increased P3b amplitude during a counting oddball task, but not during its motor counterpart (Morand-Beaulieu et al., 2016). It was thus hypothesized that motor-related ERPs associated with motor responses could mask the treatment effects that were seen in the counting oddball task. Another explanation could be that both versions of the task differ in terms of test-retest reliability, thus affecting the capacity to detect treatment effects on ERPs. Understanding whether test-retest reliability differs across these two tasks is therefore relevant for such experiments involving repeated assessments.

Non-clinical studies have separately reported good test-retest reliability for motor (Hall et al., 2006; Sandman & Patterson, 2000; Segalowitz & Barnes, 1993; Williams et al., 2005) and counting oddball tasks (Walhovd & Fjell, 2002), as well as oddball tasks combining motor responses and silent count of deviant stimuli (Kinoshita et al., 1996; Maeda et al., 1995). However, no study has ever compared the test-retest reliability of these two variants of the oddball task in a head-to-head comparison.

To date, only a few studies performed such assessment of test-retest reliability of ERPs across various tasks within the same group of individuals. A study assessed the P3b across modalities in visual and auditory oddball tasks and reported good test-retest reliability for both procedures (Mathalon et al., 2000). Another study found that the P100 and N100 peak amplitudes elicited by a motor oddball task and a Sternberg task showed good and similar 4-month test-retest reliability (Cassidy et al., 2012). The P100 and N100 are two ERPs reflecting early attentional processes (Luck, 2005). In tasks using visual stimuli, they are mostly generated by the middle occipital and fusiform gyri (Herrmann & Knight, 2001; Martínez et al., 1999).

Here, we decided to add another experimental paradigm to our multi-task assessment: the stimulus-response compatibility (SRC) paradigm. This type of task assesses cognitive control by presenting stimuli whose position or orientation is either compatible or incompatible with the response to be given. Reaction times are generally longer when a stimulus' attributes are incompatible with the required response: a phenomenon called the Simon effect (Simon & Wolf, 1963). SRC paradigms elicit an N200 and a P3b, but they also allow to study lateralized readiness potentials (LRP), which are associated with movement preparation (Luck, 2005). In this type of task, the N200 indexes conflict monitoring and inhibitory processes (Folstein & Van Petten, 2008) and is mostly generated by the anterior cingulate cortex (Huster et al., 2010; Parvaz et al., 2014). In other tasks involving cognitive control such as the Go/No-Go task and the Continuous Performance Test (CPT), where participants must press a button for one or several types of stimuli but inhibit their response when a specific stimulus is presented (Conners et al., 2018), the N200 and the P3b were found to have moderate to excellent test-retest reliability (Brunner et al., 2013; Fallgatter et al., 2002; Hammerer et al., 2013; Segalowitz et al., 2010). ERPs elicited by SRC paradigms offer a similar evaluation of the brain correlates of cognitive control but their test-retest reliability remains to be established.

Measuring ERPs in three different paradigms also allows to assessing how ERPs are correlated across tasks. To our knowledge, there have been very few assessments of correlations between ERPs elicited by different cognitive tasks. For instance, Riesel et al. (2013) reported that error-related components measured in Flanker, Stroop, and Go/No-Go tasks showed high correlations. Our dataset has the potential to further improve the knowledge about cross-task correlations of ERPs.

Therefore, the current study aimed to fill some gaps in the literature. By comparing ERPs in two different visual oddball tasks, our first aim was to assess whether motor responses impacted their test-retest reliability. It was hypothesized that test-retest reliability would not differ between both versions of the oddball task. Secondly, we also aimed to assess the test-retest reliability of several ERPs elicited by the SRC task. According to previous studies of test-retest reliability in analogous cognitive control tasks, we hypothesized that ERPs elicited by the SRC task would show good test-retest reliability. A third goal was to evaluate how ERPs across the three tasks were correlated. We hypothesized that ERPs would be correlated across tasks, especially for both variants of the oddball task. Finally, this study aimed to assess whether motor responses contribute to the P3b amplitude. It was expected that the P3b peak would be larger in the counting than the motor oddball task.

Methods

Participants

Thirty-five right-handed healthy participants were recruited to take part in this project. Inclusion criteria were: (i) right-hand dominant; (ii) age 18–65 years; and (iii) good or corrected vision and normal color perception. Exclusion criteria consisted of: (i) history of neurological or psychiatric disorder; (ii) presence of head injury in the last year; (iii) psychiatric medication uptake; and (iv) misuse of alcohol or drugs. After the first assessment (T1), one participant was excluded because of psychiatric medication uptake and another was excluded because of impaired color perception. Seven other participants dropped out and did not

return for the second assessment. Therefore, the final group comprised of 26 participants[1] (10 females, M_{age} = 37 years, *SD* = 11.3) who returned to the laboratory for the second assessment (T2) after approximately 4 months (*M* = 133.3 days, *SD* = 23.4 days). Their socio-demographic characteristics are displayed in Table 1. This study was approved by the ethics committee of the Institut universitaire en santé mentale de Montréal (#2012-029) and informed consent was obtained from all participants prior to their participation in the study.

Table 1. Sociodemographic data

	Mean	*SD*
Age	37	11.3
Sex (M:W)	16:10	N/A
Intelligence (RPM percentile)	78	22.1
Handedness (R:L)	26:0	N/A
EHI score	88.7%	16.3%
Color perception (Ishihara)	10.5	0.6
Visual acuity (Snellen)	85.8%	22.0%
Anxiety (BAI)	3	3.9
Depression (BDI)	3	4.0

Note. *SD* = Standard Deviation; BDI = Beck Depression Inventory; BAI = Beck Anxiety Inventory; EHI = Edinburgh Handedness Inventory; RPM = Raven's Progressive Matrices.

Procedure

Before the EEG recording at the first assessment, depression and anxiety symptoms were respectively assessed with the Beck Depression Inventory (BDI; Beck et al., 1961) and the Beck Anxiety Inventory (BAI; Beck et al., 1988). General intelligence was assessed with the Raven's Progressive Matrices (RPM; Raven, 1938). The RPM is a test of nonverbal reasoning and consists of 60 geometrics patterns presented as matrices with a missing piece. Participants are presented with several response choices and must select the right one to fill in the missing piece. Also, visual acuity and color perception were respectively assessed with the Snellen Chart (Snellen, 1862) and the Ishihara test (Ishihara, 1917).

Participants were then seated in a dimly lit room in a front of a computer screen (Viewsonic SVGA 17″ monitor, 1,280 × 1,024 resolution) on which stimuli were presented. They first performed an oddball task, followed by the SRC task, and then by another oddball task. The order of presentation of the counting and the motor tasks was counterbalanced across participants and assessments.

Stimulus-Response Compatibility Task

Left- and right-pointing colored arrows (186 × 150 pixels) were randomly presented on a white background, during a single block. Arrows were presented for 200 ms, with an interstimulus interval (fixation cross) randomly ranging from 1,500 to 1,800 ms. Participants responded on a keyboard, by pressing either the left arrow key with the left index finger or the right arrow key with the right index finger. In the compatible condition (100 blue arrows; RGB (0, 0, 255)), participants pressed the arrow key corresponding to the direction of the arrow presented on the computer screen. In the incompatible condition (100 black arrows; RGB (0, 0, 0)), participants pressed the arrow key corresponding to the opposite direction of the arrow presented on the computer screen. In the No-Go condition (50 red arrows; RGB (255, 0, 0)), participants were instructed to give no response. Arrows' direction was equally distributed across the condition. The response window started with stimulus onset and ended with the offset of the fixation cross.

Visual Counting Oddball Task

In this task, black (RGB (0, 0, 0)) Arial letters (X and O; 19 × 21 pixels) were randomly presented on a white background during a single block. Stimulus duration was 100 ms, with an interstimulus interval (fixation cross) randomly ranging from 1,500 to 1,700 ms. The letters "O" and "X" were respectively the standard and deviant stimuli. There were 200 trials in total: 160 standard trials and 40 deviant trials. In this task, participants were asked to count the number of deviant stimuli and to report the exact number at the end of the task.

Visual Oddball Task With Motor Responses

This task used the same stimuli and parameters as the visual counting oddball task, but participants were asked to press the left arrow key with their left index finger when standard stimuli were presented and to press the right arrow key with their right index finger when deviant stimuli were presented. The response window started with stimulus onset and ended with the offset of the fixation cross.

Electrophysiological Recordings and Signal Extraction

The electroencephalogram (EEG) was recorded while participants performed the three experimental tasks, at both the first and the second assessments. These assessments were separated by a 4-month interval. The EEG

[1] These 26 participants constituted the control group in an earlier publication from our research group, where their EEG data collected during the SRC task were compared to that of a group with Tourette syndrome (Morand-Beaulieu et al., 2018). The EEG data collected during both oddball tasks have not been published elsewhere.

signals were recorded from 62 Ag/AgCl electrodes mounted in a lycra cap (Electrode Arrays, El Paso, TX, USA), placed according to the standard EEG guidelines (American EEG Society, 1994), and referenced to the nose. EEG was recorded through IWave (InstEP Systems, Montreal, QC, Canada) with a digital amplifier (Sensorium Inc., Charlotte, VT, USA). EEG was sampled continuously at 500 Hz and recorded with an analog high-pass filter of 0.01 Hz and a low-pass filter of 100 Hz. Impedance was kept below 5 KΩ with an electrolyte gel (JNetDirect Biosciences, Herndon, VA). Additional electrodes were placed at the outer canthus of each eye and below and above the left eye to correct ocular artifacts. Stimuli presentation was monitored by presentation (Neurobehavioral Systems, Albany, CA, USA).

Raw EEG signals from each task were corrected offline for ocular artifacts with the Gratton algorithm, which is a regression-based method using the duration and amplitude of eyeblinks to correct eyeblink artifacts (Gratton et al., 1983). EEG was averaged time-locked to stimulus onset (and to response onset in the SRC task). For both the SRC and the motor oddball tasks, only trials with correct responses were included in the averaged ERP. Averaged data were digitally filtered with a 101-tap FIR filter designed with a Hamming window (0.3–30 Hz bandpass). The minimum stopband attenuation was 54 dB and the transition region width was 16.5 Hz. Clippings due to amplifier saturation and remaining epochs exceeding 100 μV were removed. In the SRC task, approximately 6% of the trials were excluded because of incorrect answers or artifacts (other than ocular artifacts). In the motor and counting oddball tasks, 9% and 4% of the trials were respectively removed (see Electronic Supplementary Material, ESM 1, Table E1 for the number of trials per condition).

In both oddball tasks, the following ERPs were assessed: P100 (60–110 ms), N100 (110–180 ms), P200 (130–240 ms), N200 (180–300 ms), and P3b (250–550 ms). The difference waveform (deviants minus standards) was also calculated and the difference waveform P3b was assessed in the same time window. In the SRC task, the N200 (150–300 ms) and P3b (250–550 ms) were assessed. For each ERP in each task, the maximum peak and its latency were assessed. Following Cassidy et al. (2012), the mean amplitude and the area under the curve of the P3b were also assessed during both oddball tasks. In order to limit the number of comparisons, ERPs were assessed at midline electrodes where the grand average mean of T1 and T2 amplitudes was maximal, except for the N100 and P100 which were assessed at lateral parieto-occipital electrodes (PO7/PO8) (Cassidy et al., 2012). Electrodes used for analyses are depicted in Figure 1.

The LRPs were computed through a double subtraction of electrodes C3 and C4 as proposed by Coles (1989):

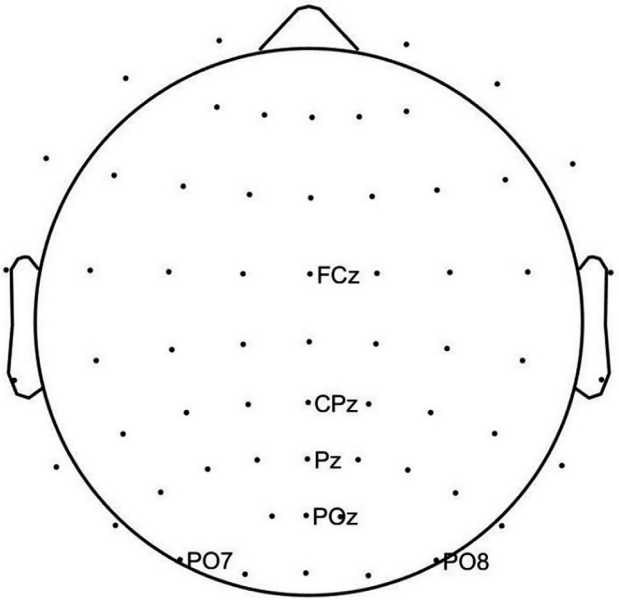

Figure 1. Layout of the 62-channel EEG cap. Electrodes that were used to assess test-retest reliability are shown with their label.

$$\mathrm{LRP} = \frac{\left[\mathrm{Mean}\,(\mathrm{C4}-\mathrm{C3})^{\text{left hand}} + \mathrm{Mean}\,(\mathrm{C3}-\mathrm{C4})^{\text{right hand}}\right]}{2}. \quad (1)$$

LRP peaks and onsets were measured from 150 to 900 ms relative to stimulus onset for stimulus-locked LRP (sLRP) and from −500 to 0 ms relative to response onset for response-locked LRP (rLRP). sLRP and rLRP onsets were calculated with the relative criterion method (Smulders et al., 1996), which was set at 20%.

For ERP and LRP analyses, we only included participants for which we could reliably identify a peak in the time window of interest. Therefore, for some components presented in Tables 2, 3, and 4, the included n is smaller than 26. Scalp topographies were computed in EEGLAB (Delorme & Makeig, 2004), using the *topoplot* function with the "maxmin" parameter for map limits. Thus, colors on individual scalp topographies represent the lower and upper amplitude limits at a given time point and do not represent the same amplitude value across scalp topographies.

Statistical Analyses

Task performance (RT and accuracy) was assessed with one-way analysis of variance (ANOVAs), with two within-subjects factors: Time (T1/T2) and Condition (SRC task: compatible/incompatible/No-Go; motor oddball task: standard/deviant). A paired t-test was used to assess if the number of deviant stimuli identified by participants differed between sessions.

Table 2. Test-retest reliability of the counting oddball task

ERP	Electrode	Measure	Condition	*n*	Mean T1	*SD*	Mean T2	*SD*	*t*	*r*	*r* 95% CI	ICC	ICC 95% CI
P100	PO8	Peak amplitude	Standard	22	4.1	3.1	4.3	2.9	−0.76	.90***	.76–.96	.90***	.77–.96
			Deviant	25	4.2	3.3	4.0	3.6	0.49	.84***	.66–.92	.84***	.68–.93
		Peak latency	Standard	22	85.6	11.3	87.2	11.1	−0.74	.57**	.20–.80	.58**	.22–.80
			Deviant	25	85.3	9.2	87.8	8.7	−1.36	.45*	.05–.70	.44*	.08–.71
N100	PO7	Peak amplitude	Standard	24	−2.1	1.7	−2.6	1.8	1.76†	.75***	.49–.88	.73***	.48–.88
			Deviant	24	−6.4	3.6	−7.0	3.6	1.42	.83***	.63–.92	.82***	.64–.92
		Peak latency	Standard	24	142.7	17.5	145.7	18.8	−0.93	.62**	.28–.81	.62***	.31–.82
			Deviant	24	145.5	13.5	147.8	14.7	−1.20	.78**	.54–.90	.77***	.56–.89
P200	POz	Peak amplitude	Standard	20	4.2	3.7	4.2	3.5	−0.07	.76***	.47–.90	.77***	.50–.90
		Peak latency	Standard	20	199.0	37.4	195.4	37.0	0.31	.01*ns*	−.43–.44	.01*ns*	−.46–.45
N200	FCz	Peak amplitude	Deviant	26	−3.9	4.4	−3.6	5.8	−0.38	.69***	.41–.85	.67***	.39–.84
		Peak latency	Deviant	26	246.3	26.7	253.0	25.6	−1.23	.44*	.06–.70	.43*	.07–.70
P3b	CPz	Peak amplitude	Deviant	26	16.1	5.7	15.4	6.3	0.92	.80***	.59–.90	.80***	.61–.91
		Mean amplitude	Deviant	26	8.7	3.9	8.1	4.4	1.14	.78***	.56–.89	.77***	.55–.89
		AUC	Deviant	26	2,702.4	1,116.6	2,515.6	1,249.6	1.20	.78***	.56–.89	.77***	.56–.89
		Peak latency	Deviant	26	391.5	41.3	395.0	46.5	−0.51	.69***	.41–.85	.70***	.43–.85
		Diff peak amplitude	Deviant	26	13.6	4.0	12.7	4.5	1.35	.64***	.33–.82	.63***	.33–.81
		Diff mean amplitude	Deviant	26	6.3	2.8	5.2	3.0	1.77†	.41*	.03–.68	.39*	.04–.67
		Diff AUC	Deviant	26	2,035.6	738.2	1,746.2	772.0	2.07*	.56**	.22–.77	.53**	.19–.75
		Diff peak latency	Deviant	26	397.1	39.3	393.8	52.7	0.48	.75***	.50–.88	.72***	.47–.87

Note. AUC = area under the curve; Diff = difference; *ns* = not significant. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 3. Test-retest reliability of the motor oddball task

ERP	Electrode	Measure	Condition	*n*	Mean T1	*SD*	Mean T2	*SD*	*t*	*r*	*r* 95% CI	ICC	ICC 95% CI
P100	PO8	Peak amplitude	Standard	22	3.8	2.7	4.3	2.9	−1.49	.88***	.72–.95	.87***	.71–.94
			Deviant	23	4.8	3.5	4.9	4.0	−0.27	.80***	.57–.91	.80***	.58–.91
		Peak latency	Standard	22	85.4	8.3	87.6	9.5	−1.56	.71***	.40–.87	.69***	.40–.86
			Deviant	23	85.1	9.0	88.0	8.8	−1.24	.23*ns*	−.20–.58	.22*ns*	−.19–.57
N100	PO7	Peak amplitude	Standard	23	−3.6	2.5	−3.7	2.3	0.27	.69***	.38–.85	.70***	.41–.86
			Deviant	26	−7.1	5.2	−7.8	4.6	1.09	.79***	.57–.90	.78***	.57–.90
		Peak latency	Standard	23	145.8	17.2	143.8	17.3	0.63	.61**	.26–.81	.62**	.28–.82
			Deviant	26	140.5	10.0	145.1	12.9	−2.61*	.72***	.45–.86	.65***	.33–.83
P200	POz	Peak amplitude	Standard	19	5.0	4.7	5.3	3.8	−0.38	.79***	.51–.91	.78***	.51–.91
		Peak latency	Standard	19	215.6	16.6	209.1	29.5	0.82	−.06*ns*	−.48–.40	−.05*ns*	−.50–.41
N200	FCz	Peak amplitude	Deviant	26	−3.8	5.6	−5.6	5.7	2.35*	.78***	.56–.89	.75***	.49–.88
		Peak latency	Deviant	26	237.5	22.1	241.6	23.9	−0.89	.48*	.11–.72	.48**	.13–.73
P3b	CPz	Peak amplitude	Deviant	26	18.1	8.7	17.4	7.7	0.61	.75***	.50–.88	.74***	.51–.88
		Mean amplitude	Deviant	26	10.2	6.3	9.6	5.6	0.78	.80***	.59–.90	.80***	.60–.90
		AUC	Deviant	26	3,129.9	1,820.6	2,978.8	1,588.7	0.70	.80***	.59–.90	.79***	.59–.90
		Peak latency	Deviant	26	380.9	54.2	394.5	60.5	−1.82†	.78***	.56–.89	.76***	.53–.89
		Diff peak amplitude	Deviant	26	10.6	5.4	10.7	5.8	−0.23	.72***	.45–.86	.72***	.47–.87
		Diff mean amplitude	Deviant	26	4.8	3.7	4.5	4.0	0.56	.70***	.42–.85	.71***	.45–.86
		Diff AUC	Deviant	26	1,570.2	1,044.1	1,545.5	1,092.4	0.15	.71***	.44–.86	.71***	.45–.86
		Diff peak latency	Deviant	26	391.1	61.1	407.4	49.5	−1.43	.46*	.09–.71	.44**	.09–.70

Note. AUC = area under the curve; Diff = difference; *ns* = not significant. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Across the different ERPs, tasks, and conditions, 66 ERP metrics were assessed at both T1 and T2, for a total of 132. The normality of data distribution was assessed with the Kolmogorov-Smirnov (K-S) test. Given that the K-S test did not reach the significance threshold for almost every ERP metric (94%), test-retest reliability was assessed with

Table 4. Test-retest reliability of the SRC task

ERP	Electrode	Measure	Condition	*n*	Mean T1	*SD*	Mean T2	*SD*	*t*	*r*	*r* 95% CI	ICC	ICC 95% CI
N200	FCz	Peak amplitude	Comp	26	−3.1	3.3	−4.2	3.7	−1.31	.76***	.52–.88	.73***	.47–.87
			Incomp	26	−3.4	4.0	−4.0	3.9	1.24	.83***	.64–.92	.83***	.66–.92
			No-Go	26	−3.3	3.6	−4.1	3.3	1.17	.70***	.42–.85	.68***	.41–.84
		Peak latency	Comp	26	209.1	28.7	215.6	29.2	2.24*	.62**	.30–.81	.61***	.31–.80
			Incomp	26	222.8	22.7	217.2	26.6	1.33	.57**	.23–.78	.56**	.23–.77
			No-Go	26	216.5	22.6	211.3	25.5	1.65	.58**	.24–.78	.57**	.25–.78
P3b	Pz	Peak amplitude	Comp	26	10.2	5.8	10.2	5.5	−1.12	.88***	.74–.94	.88***	.75–.94
	Pz		Incomp	26	9.4	5.8	10.0	5.8	−0.78	.86***	.70–.93	.86***	.71–.93
	FCz		No-Go	26	12.7	5.2	12.3	5.7	0.02	.85***	.68–.93	.85***	.70–.93
	Pz	Mean amplitude	Comp	26	5.0	3.9	5.3	4.5	−1.23	.86***	.70–.93	.85***	.70–.93
	Pz		Incomp	26	4.4	4.8	4.8	4.6	−0.94	.88***	.74–.94	.88***	.75–.94
	FCz		No-Go	26	6.1	4.0	5.7	4.5	0.75	.81***	.61–.91	.81***	.61–.91
	Pz	AUC	Comp	26	1,578.3	1,108.5	1,715.1	1,237.1	0.01	.87***	.72–.94	.86***	.71–.93
	Pz		Incomp	26	1,444.4	1,338.4	1,584.2	1,257.9	−1.12	.91***	.80–.96	.90***	.80–.96
	FCz		No-Go	26	1,940.1	1,146.3	1,860.6	1,246.5	0.70	.84***	.66–.92	.84***	.67–.92
	Pz	Peak latency	Comp	26	378.9	65.9	394.6	60.7	−0.93	.36†	−.03–.65	.36*	−.02–.65
	Pz		Incomp	26	375.4	74.5	392.5	67.8	−1.26	.50**	.14–.74	.50**	.15–.74
	FCz		No-Go	26	419.8	65.0	419.6	52.3	0.58	.72***	.45–.86	.73***	.48–.87
sLRP	C3'	Peak amplitude	Comp	23	−2.2	0.9	−2.6	1.1	2.09*	.48*	.08–.74	.44*	.07–.71
			Incomp	23	−2.6	1.2	−2.9	1.0	0.02	.66**	.33–.84	.64***	.33–.83
		Onset latency	Comp	23	230.2	60.4	230.0	55.2	1.32	.30^{ns}	−.13–.63	.31†	−.13–.64
			Incomp	23	333.5	58.2	321.9	47.3	1.23	.65**	.32–.83	.64***	.32–.83
rLRP	C3'	Peak amplitude	Comp	23	−3.0	1.1	−3.3	1.2	1.26	.59**	.23–.80	.58**	.24–.79
			Incomp	23	−3.2	1.3	−3.8	1.8	2.51*	.73***	.45–.87	.65***	.30–.84
		Onset latency	Comp	23	−233.4	54.7	−245.9	44.1	0.96	.22^{ns}	−.21–.51	.22^{ns}	−.20–.57
			Incomp	23	−252.8	52.2	−247.0	47.0	−0.73	.71***	.41–.86	.71***	.44–.87

Note. The shaded areas highlight the most important data. AUC = area under the curve; Comp = compatible; Diff = difference; Incomp = incompatible; sLRP = stimulus-locked lateralized readiness potentials; rLRP = response-locked lateralized readiness potentials; SRC = stimulus-response compatibility; *ns* = not significant. $^{\dagger}p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Pearson's correlations and intraclass correlation coefficients (ICC) to ensure consistency with previous reports of test-retest reliability of ERPs (Cassidy et al., 2012; Kinoshita et al., 1996; Maeda et al., 1995; Munsters et al., 2019). The ICC is a measure that accounts for both the consistency of a participant's data and the change in the average group data across assessments (Vaz et al., 2013). The following settings were used in ICC analyses: two-way mixed model, absolute agreement, single measures (Koo & Li, 2016; Munsters et al., 2019). Thus, ICCs and Pearson's correlations were used as test-retest reliability coefficients in the current study. They were interpreted as follows: very weak (.00–.19), weak (.20–.39), moderate (.40–.59), strong (.60–.79), and very strong (.80–1.00) (Evans, 1996; Landis & Koch, 1977).

To compare the reliability coefficients from the motor and the counting oddball tasks, we used the procedure of Hittner et al. (2004), which is a modification of Dunn and Clark's (1971) *z* using a back-transformed average Fisher (1921) Z procedure. This test is implemented in the *cocor* (http://comparingcorrelations.org) online software (Diedenhofen & Musch, 2015). Correlations were also performed between ERP components elicited by the three tasks. Finally, to assess whether P3b amplitude differed across oddball task variants, we performed a repeated-measures ANOVA with the within-subjects factor Time (T1/T2) and Variant (Motor/Counting).

Results

Behavioral Data

In the counting oddball task, the number of deviant stimuli identified by participants did not differ between assessments, $t(25) = -0.95$, $p = .35$, $d = .21$ (Figure 2A). In the motor oddball task, there was a better accuracy, $F(1, 25) = 43.10$, $p < .001$, $d = 1.47$, as well as faster RT, $F(1, 25) = 55.01$, $p < .001$, $d = .87$, for standard than for deviant stimuli. However, there was no difference between assessments for both accuracy, $F(1, 25) = 2.66$, $p = .12$, $d = .30$, and RT, $F(1, 25) = 1.26$, $p = .27$, $d = .18$, (Figure 2B).

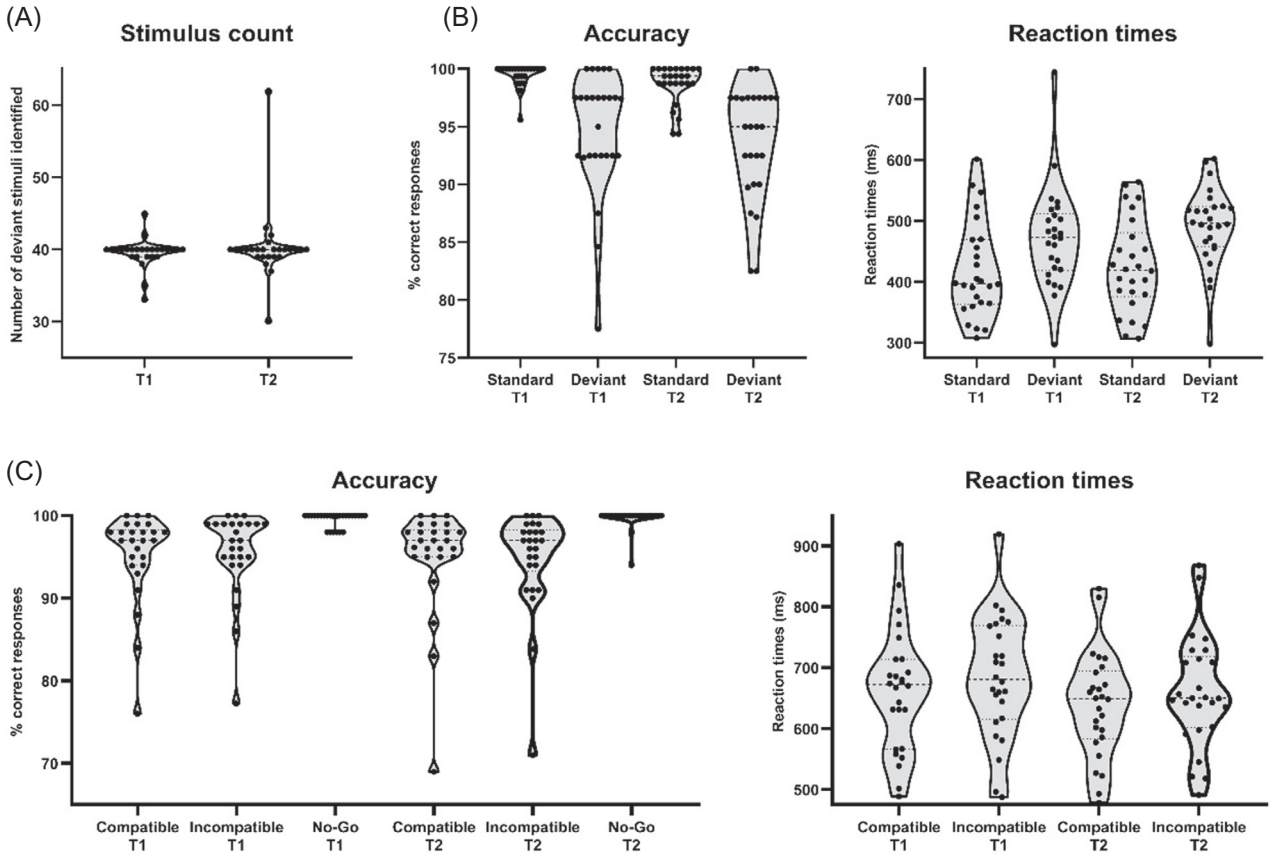

Figure 2. Behavioral performance. This figure shows violin plots of participants' behavioral performance in each of the three experimental tasks. (A) In the counting oddball task, the number of deviant stimuli identified did not differ across assessments. (B) In the motor oddball task, participants showed better accuracy and faster RT for standard than deviant stimuli. (C) In the SRC task, participants showed better accuracy for No-Go than compatible and incompatible stimuli. They also provided faster responses at T2 compared to T1.

In the SRC task, RT was slightly faster during the compatible trials, $F(1, 25) = 15.74$, $p = .001$, $d = .23$. Participants were also slightly faster at the second assessment, $F(1, 25) = 6.37$, $p = .018$, $d = .27$. There was a compatibility effect regarding accuracy, $F(2, 50) = 13.59$, $p < .001$, $\eta_p^2 = .352$. A Bonferroni post hoc test revealed that accuracy was better in the No-Go than in the compatible and incompatible conditions. There was however no significant difference regarding accuracy across assessments, $F(1, 25) = 0.37$, $p = .55$, $d = .08$ (Figure 2C). Correlations between behavioral measures are reported in ESM 1.

Test-Retest Reliability of ERP Components

Mean peak latency and amplitude as well as test-retest reliability coefficients for the counting oddball task, motor oddball task, and SRC task are respectively presented in Tables 2, 3, and 4. In both oddball tasks, the P100 and the N100 were elicited by standard and deviant stimuli. The P200 was elicited by standard stimuli and the N200 and the P3b were elicited by deviant stimuli. Scalp topographies and butterfly plots depicting grand average ERPs during the counting and the motor oddball tasks are presented in Figure 3 and Figure 4, respectively.

In the counting oddball task, the peak amplitude of the P100, the deviant N100, and the P3b fell in the strong to the very strong range, while the standard N100, the N200, and P200 peak amplitudes showed moderate to very strong reliability. Globally, peak latency measures seemed to be somewhat less reliable than peak amplitude measures. The latencies of the deviant N100 and the P3b showed moderate to very strong reliability. However, for other ERPs, the lower bound of the confidence interval fell into the weak or very weak category.

The motor oddball task yielded a similar picture, where the P100 amplitude showed strong to very strong test-retest reliability, while the amplitude of the other ERPs showed moderate to very strong reliability. Similar to the counting oddball task, the latency was slightly less reliable. The standard P100 and the P3b latencies fell in the moderate

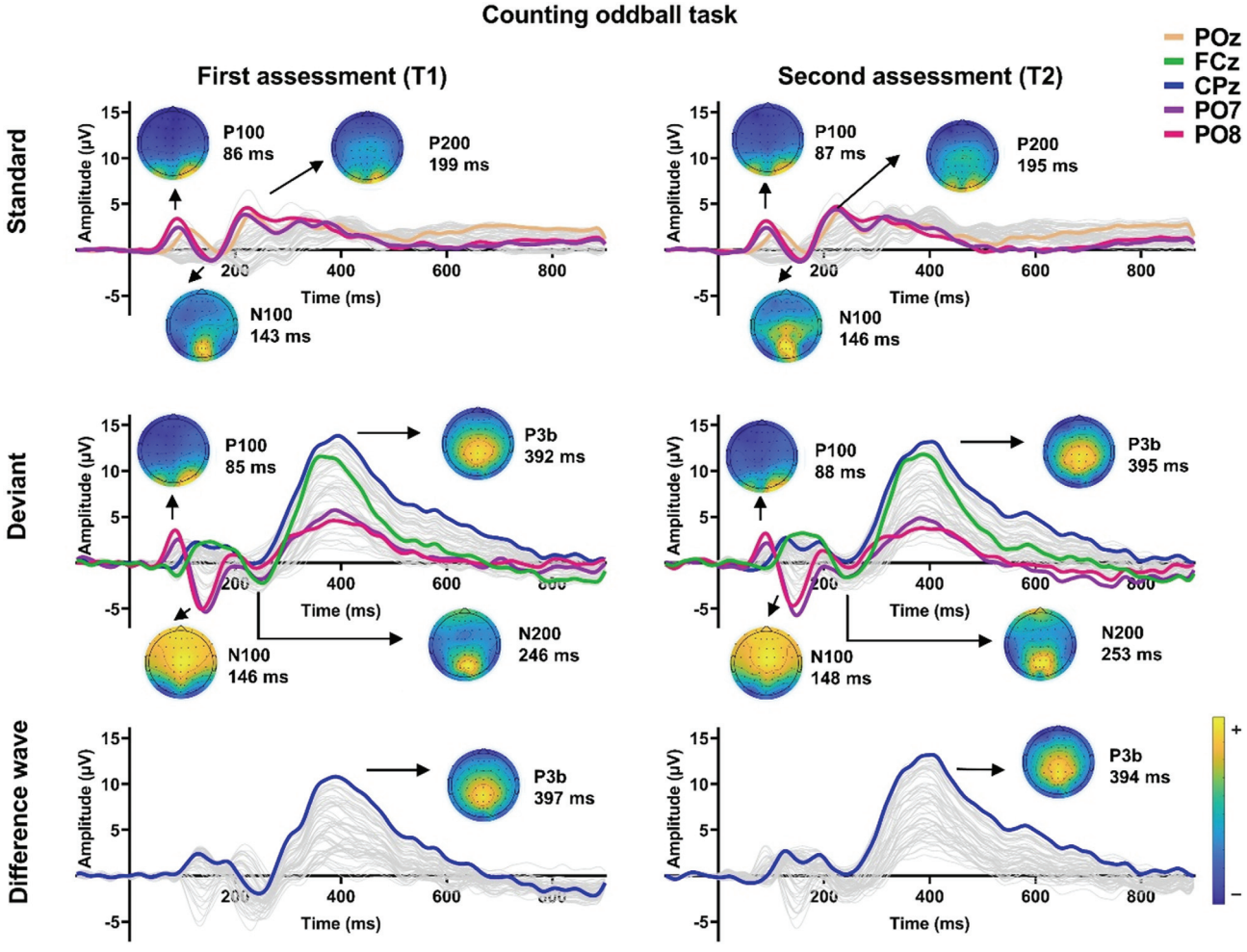

Figure 3. ERP waveforms during the counting oddball task. This figure presents the butterfly plots of the grand average ERP during the counting oddball task, for each assessment and each condition as well as the difference between the deviant and standard conditions. All 62 channels are depicted in light gray, excepted for the colored channels that were used to assess the following ERPs: P100 (PO8 – pink), N100 (PO7 – purple), P200 (POz – beige), N200 (FCz – green), and P3b (CPz – blue). Scalp topographies corresponding to the peak latency of each ERP are also depicted.

to the strong range, while the N100 latency showed weak to very strong reliability. For the deviant N100, P200, and N200 latencies, the lower bound of the confidence interval fell into the very weak category.

In the SRC task, the N200 and the P3b were elicited by the compatible, incompatible, and No-Go conditions. In this task, the P3b peak amplitude, mean amplitude, and area under the curve all showed strong to very strong test-retest reliability in each condition. However, the P3b latency was less reliable. Only the latency of the No-Go P3b showed moderate to very strong reliability. Regarding the N200 peak amplitude, test-retest reliability coefficients were in the moderate to the very strong range. However, the latency of the N200 seemed somewhat less reliable. ERP waveforms for the SRC task are presented in Figure 5.

Regarding LRPs elicited during the SRC task, reliability coefficients for peak amplitude and incompatible onset suggested moderate to strong reliability, but the confidence intervals were very large, which calls for caution regarding their interpretation. LRP waveforms are depicted in Figure 6.

Comparison of Reliability Coefficients Across Both Oddball Tasks

Test-retest reliability did not differ between the counting and the motor oddball task. There was only a trend-level difference suggesting greater test-retest reliability of the P3b difference waveform peak latency in the counting oddball task, compared to the motor oddball task, $z = 1.90$,

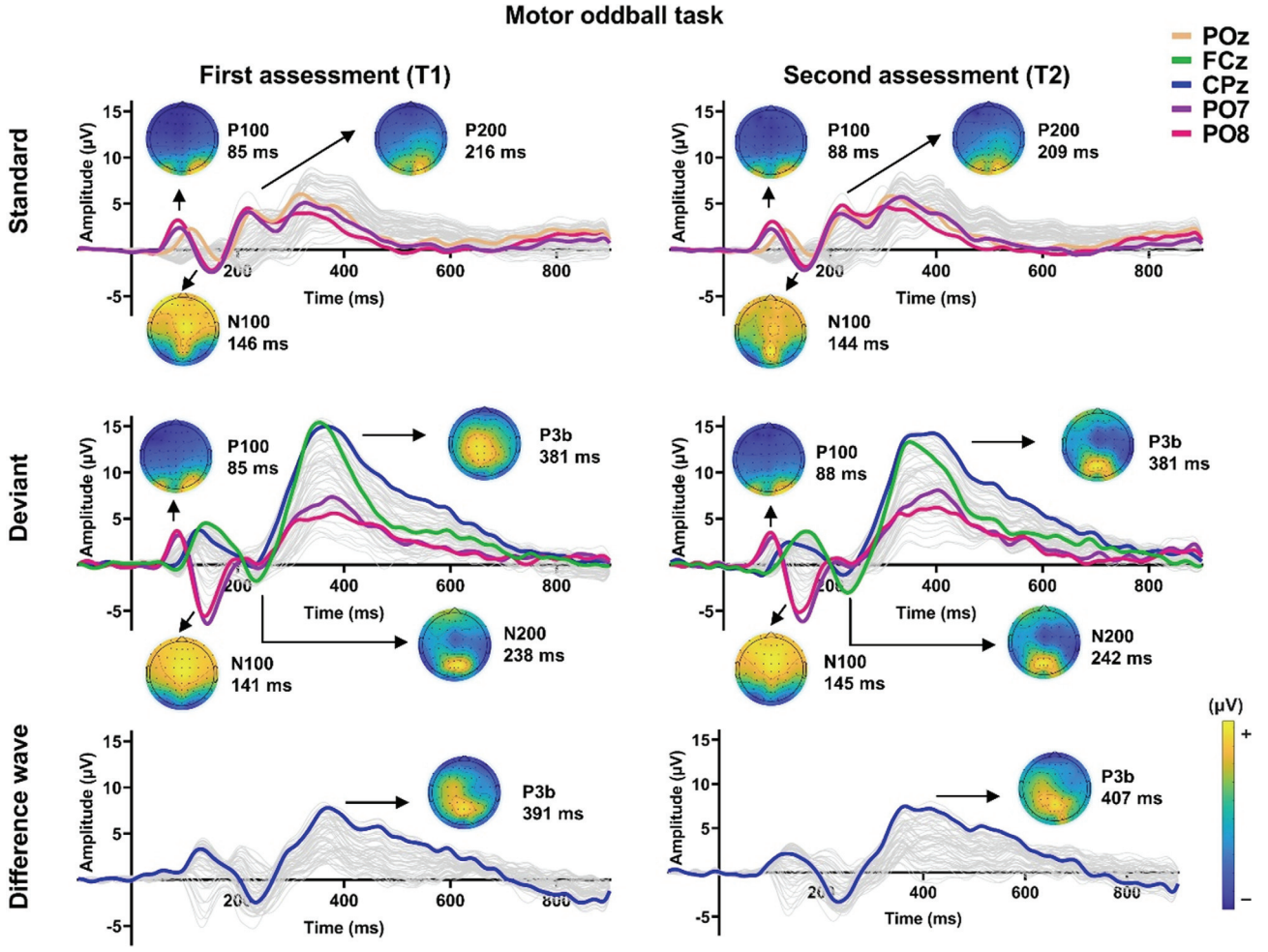

Figure 4. ERP waveforms during the motor oddball task. This figure presents the butterfly plots of the grand average ERP during the motor oddball task, for each assessment and each condition as well as the difference between the deviant and standard conditions. All 62 channels are depicted in light gray, excepted for the colored channels that were used to assess the following ERPs: P100 (PO8 – pink), N100 (PO7 – purple), P200 (POz – beige), N200 (FCz – green), and P3b (CPz – blue). Scalp topographies corresponding to the peak latency of each ERP are also depicted.

p = .058. All other comparisons between test-retest reliability coefficients from both tasks were nonsignificant, all p-values > .1, suggesting comparable test-retest reliability in the counting and motor oddball tasks.

Correlations Between Tasks

The full correlation matrix of all variables in the study is shown in ESM 2, Table E2. Tables 5, 6, 7, 8, and 9 present correlations between the P100, the N100, the P200, the N200, and the P3b across tasks. The N100 and P100 peak amplitudes were highly correlated between both oddball tasks. Latencies of both components were also correlated, except for the standard P100 at the first assessment. The P200 peak amplitude and latency measured in both oddball tasks were also correlated, though this correlation only reached trend-level for the P200 latency at the first assessment. The N200 peak amplitude was correlated between all tasks. However, the N200 was not. The deviant P3b elicited by the motor and the counting oddball tasks were also correlated. Globally, the P3b (peak amplitude, mean amplitude, area under the curve) measured in oddball tasks was correlated to the P3b measured in the SRC task. However, this was not the case for P3b latency.

Comparison of P3b Amplitude Between Oddball Tasks

The P3b peak was larger in the motor than in the counting oddball task, $F(1, 25) = 5.50$, $p = .027$. There was no main effect or interaction involving the repeated assessment factor.

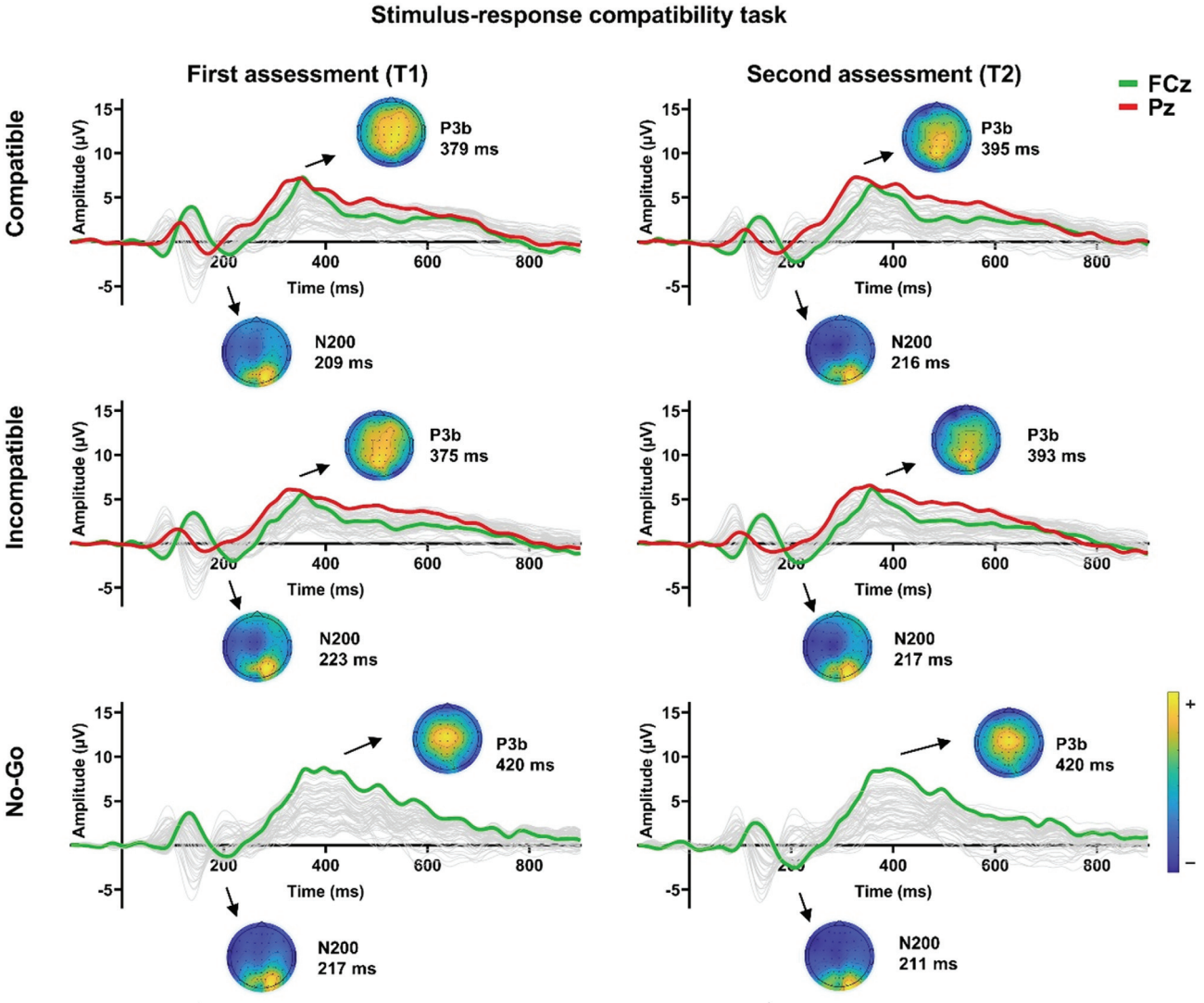

Figure 5. ERP waveforms during the stimulus-response compatibility task. This figure presents the butterfly plots of the grand average ERP during the stimulus-response compatibility task, for each assessment and each condition. All 62 channels are depicted in light gray, excepted for the colored channels that were used to assess the following ERPs: N200 (FCz – green), and P3b (Pz – red). Given the anteriorization of the P3b during No-Go trials, the No-Go P3b was assessed at electrode FCz. Scalp topographies corresponding to the peak latency of each ERP are also depicted.

Discussion

The main goal of this study was to compare the test-retest reliability across counting and motor oddball tasks. We also aimed at validating the test-retest reliability of the SRC paradigm. Overall, our analyses revealed moderate to very strong test-retest reliability for most ERP components. Pearson's correlations and intraclass correlations yielded very similar reliability coefficients. Also, we aimed to look at correlations between ERPs elicited by different tasks and found that the peak amplitudes for the P100, N100, N200, P200, and P3b were strongly correlated across tasks.

In both oddball tasks, all components (P100, N100, N200, P200, and P3b) showed the expected topographical distribution, as can be seen in Figure 3 and Figure 4. The P100, N100, and P200 were predominantly distributed over parieto-occipital electrodes, the N200 was maximal over frontal electrodes, and the P3b had a centro-parietal topography. Our results regarding the test-retest reliability of peak amplitude measures in oddball paradigms appear to be in line with previous investigations, were moderate to very strong reliability was reported for both counting (Walhovd & Fjell, 2002) and motor (Cassidy et al., 2012; Kinoshita et al., 1996; Sandman & Patterson, 2000; Segalowitz & Barnes, 1993; Williams et al., 2005) oddball

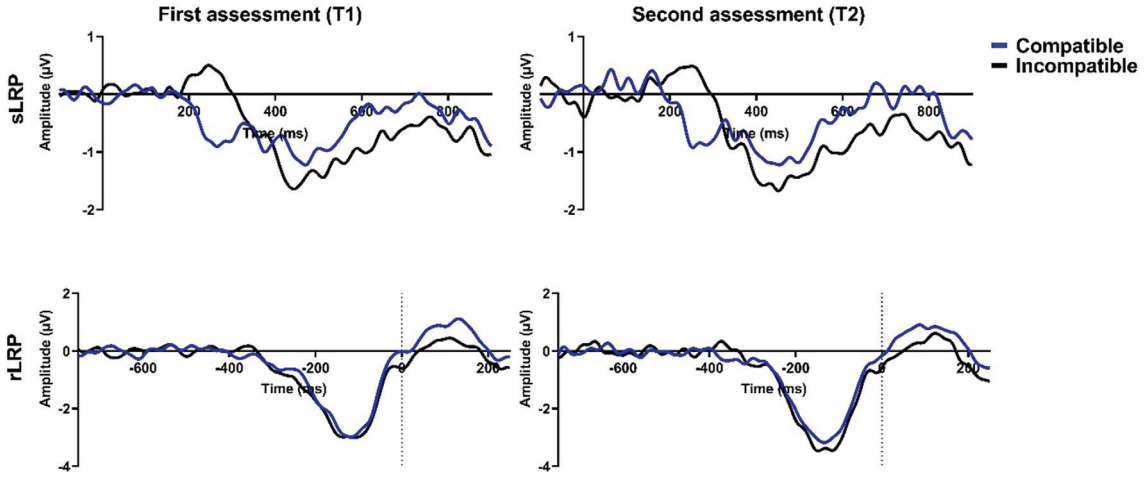

Figure 6. LRP waveforms during the stimulus-response compatibility task. This figure shows the grand average stimulus- and response-locked LRPs for each assessment and condition.

Table 5. Correlations between the P100 and N100 elicited by motor and counting oddball tasks

	Motor oddball			
	Peak amplitude		Peak latency	
Counting oddball	Standard	Deviant	Standard	Deviant
		P100 T1		
Peak amplitude				
Standard	.92***	.89***	.56**	.51*
Deviant	.81***	.81***	.58**	.47*
Peak latency				
Standard	.56**	.44*	.29	.26
Deviant	.57**	.49*	.81***	.77***
		P100 T2		
Peak amplitude				
Standard	.92***	.83***	.45*	.48*
Deviant	.87***	.87***	.16	.16
Peak latency				
Standard	.60**	.40†	.85***	.54*
Deviant	.54*	.42*	.57**	.46*
		N100 T1		
Peak amplitude				
Standard	.77***	.32	−.06	−.29
Deviant	.71***	.84***	.01	.01
Peak latency				
Standard	−.35	−.25	.63**	.40†
Deviant	−.02	.07	.02	.48*
		N100 T2		
Peak amplitude				
Standard	.68***	.48*	.10	−.32
Deviant	.61**	.93***	.08	−.07
Peak latency				
Standard	−.16	−.24	.72***	.75***
Deviant	−.10	−.10	.69***	.75***

Note. The shaded areas highlight the most important data. T1 = first assessment; T2 = second assessment. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 6. Correlations between the P200 elicited by motor and counting oddball tasks

	Motor oddball	
	Peak amplitude	Peak latency
Counting oddball	Standard	Standard
	T1	
Peak amplitude		
Standard	.68**	.44†
Peak latency		
Standard	.07	.40†
	T2	
Peak amplitude		
Standard	.50*	−.08
Peak latency		
Standard	.20	.49*

Note. The shaded areas highlight the most important data. T1 = first assessment; T2 = second assessment. †$p < .1$; *$p < .05$; **$p < .01$.

tasks. Peak latency measures were somewhat less reliable than peak amplitude measures, a finding that is also consistent with prior studies (Kinoshita et al., 1996; Walhovd & Fjell, 2002; Williams et al., 2005). Most importantly, we confirmed our hypothesis that the motor and counting oddball tasks would not differ in regard to their test-retest reliability. Therefore, while both paradigms may slightly differ in terms of cognitive demands, our results suggest that they both should provide reliable ERPs in studies using repeated assessments such as longitudinal or clinical studies.

In the SRC task, the topographical distribution of the N200 and the P3b was similar to what we observed in both oddball tasks. The N200 was maximal over fronto-central electrodes and the P3b was mostly distributed over centro-parietal electrodes. However, in the NoGo condition, we observed a shift of the P3b toward fronto-central electrodes, which is a typical pattern in Go/NoGo tasks (Fallgatter et al., 1997; Fallgatter & Strik, 1999). Test-retest reliability for peak amplitude measures was in the moderate-to-strong range for the N200 and strong-to-very-strong range for the P3b. While this is the first report of test-retest reliability of ERPs elicited through this task, our results are consistent with prior studies showing good reliability of peak amplitude measures in similar tasks involving cognitive control (Brunner et al., 2013; Fallgatter et al., 2002; Hammerer et al., 2013; Segalowitz et al., 2010). Analogous to the pattern observed in oddball tasks, test-retest reliability was somewhat reduced for latency measures, compared to peak

Table 7. Correlations between the N200 elicited by motor oddball, counting oddball, and stimulus-response compatibility tasks

					T1					
	Motor oddball		Counting oddball		Stimulus-response compatibility					
	Peak amplitude	Peak latency	Peak amplitude	Peak latency	Peak amplitude			Peak latency		
	Deviant	Deviant	Deviant	Deviant	Compatible	Incompatible	No-Go	Compatible	Incompatible	No-Go
Motor oddball										
Peak amplitude										
Deviant	–	–	.76***	−.17	.76***	.63**	.69***	−.50**	−.15	−.33†
Peak latency										
Deviant	–	–	−.35†	.47*	−.07	−.15	−.21	.36†	.04	.59**
Counting oddball										
Peak amplitude										
Deviant	–	–	–	–	.63**	.50*	.61**	−.46*	−.14	−.40*
Peak latency										
Deviant	–	–	–	–	.05	.02	−.15	.12	.12	.34†
					T2					
Motor oddball										
Peak amplitude										
Deviant	–	–	.73*	−.02	.76***	.54**	.60**	−.35†	−.36†	−.37†
Peak latency										
Deviant	–	–	−.10	.24	.00	−.06	−.18	.33†	.17	.29
Counting oddball										
Peak amplitude										
Deviant	–	–	–	–	.74***	.61**	.60**	−.32	−.36†	−.39†
Peak latency										
Deviant	–	–	–	–	.03	.06	.19	.07	.08	.29

Note. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 8. Correlations between the P3b elicited by motor oddball, counting oddball, and stimulus-response compatibility tasks (T1)

		Counting oddball								Stimulus-response compatibility											
		Peak amp	Mean amp	AUC	Peak lat	Diff peak amp	Diff mean amp	Diff AUC	Diff peak lat	Peak amp			Mean amp			AUC			Peak lat		
		D	D	D	D	D	D	D	D	C	I	N	C	I	N	C	I	N	C	I	N
Motor oddball																					
Peak amp	D	.71***	.55**	.67**	−.40*	.50**	.30	.38†	−.45*	.69***	.69***	.65***	.74***	.70***	.53**	.74***	.70***	.55**	.05	.21	−.37†
Mean amp	D	.71***	.59**	.60**	−.45*	.50*	.32	.40*	−.44*	.69***	.68***	.64***	.78***	.68***	.55**	.76***	.67***	.56**	.08	.24	−.28
AUC	D	.72***	.59**	.60**	−.43*	.49*	.30	.39*	−.44*	.68***	.68***	.65***	.77***	.68***	.54**	.77***	.68***	.57**	.10	.26	−.27
Peak lat	D	−.18	−.23	−.22	.66***	−.18	−.24	−.25	.56**	−.13	−.12	−.20	−.20	−.11	−.12	−.17	−.10	−.10	.26	.07	.37†
Diff peak amp	D	.32	.16	.16	−.01	.36†	.12	.13	−.32	.62**	.60**	.23	.60**	.59**	.07	.59**	.57**	.07	.12	.14	−.50*
Diff mean amp	D	.42*	.35†	.34†	−.16	.38†	.24	.24	−.32	.61**	.56**	.22	.61**	.55**	.12	.59**	.53**	.12**	.05	.14	−.39*
Diff AUC	D	.45*	.33†	.33†	−.12	.40*	.21	.22	−.31	.65***	.60**	.27	.64***	.59**	.15	.63**	.57**	.15	.09	.17	−.42*
Diff peak la	D	−.08	−.08	−.09	.54**	−.07	−.11	−.15	.59**	−.11	−.10	−.29	−.15	−.04	−.21	−.14	−.06	−.19	.24	.10	.42*
Counting oddball																					
Peak amp	D	–	–	–	–	–	–	–	–	.68***	.71***	.62**	.72***	.71***	.60**	.74***	.73***	.62**	.13	.30	−.05
Mean amp	D	–	–	–	–	–	–	–	–	.53**	.52**	.41*	.55**	.51**	.43*	.55**	.50*	.44*	.03	.30	.02
AUC	D	–	–	–	–	–	–	–	–	.55**	.54**	.46*	.57**	.54**	.47*	.58**	.54**	.49*	.09	.35†	.02
Peak lat	D	–	–	–	–	–	–	–	–	−.16	−.11	−.26	−.25	−.14	−.24	−.23	−.14	−.25	.11	−.12	.30
Diff peak amp	D	–	–	–	–	–	–	–	–	.61**	.65***	.34†	.61**	.63**	.30	.62**	.62**	.30	.26	.36†	.00
Diff mean amp	D	–	–	–	–	–	–	–	–	.32	.31	.04	.31	.31	.02	.31	.27	.02	.10	.33†	−.02
Diff AUC	D	–	–	–	–	–	–	–	–	.42*	.43*	.21	.42*	.42*	.21	.43*	.41*	.21	.18	.41*	−.04
Diff peak lat	D	–	–	–	–	–	–	–	–	−.19	−.22	−.30	−.27	−.25	−.16	−.26	−.25	−.18	−.11	−.35†	.44*

Note. AUC = Area under the curve; Diff = difference; lat = latency; amp = amplitude; C = compatible; D = deviant; I = incompatible; N = No-Go. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

Table 9. Correlations between the P3b elicited by motor oddball, counting oddball, and stimulus-response compatibility tasks (T2)

		Counting oddball								Stimulus-response compatibility											
		Peak amp	Mean amp	AUC	Peak lat	Diff peak amp	Diff mean amp	Diff AUC	Diff peak lat	Peak amp			Mean amp			AUC			Peak lat		
		D	D	D	D	D	D	D	D	C	I	N	C	I	N	C	I	N	C	I	N
Motor oddball																					
Peak amp	D	.76***	.79***	.78***	−.53**	.70***	.73***	.76***	−.31	.69***	.66***	.38†	.73***	.74***	.32	.75***	.74***	.32	−.10	−.12	−.16
Mean amp	D	.77***	.82***	.81***	−.56**	.65***	.65***	.68***	−.38†	.72***	.69***	.44*	.78***	.76***	.42*	.79***	.76***	.41	−.04	−.04	−.18
AUC	D	.78***	.82***	.81***	−.54**	.65***	.66***	.69***	−.36†	.70***	.68***	.46*	.78***	.76***	.42*	.80***	.77***	.42*	−.01	.02	−.14
Peak lat	D	−.33	−.40*	−.36†	.64***	−.52**	−.61***	−.56**	.58**	−.25	−.27	−.23	−.30	−.37	−.15	−.30	−.24	−.15	.04	.36†	.35†
Diff peak amp	D	.38†	.41*	.40*	−.25	.64***	.59**	.61**	−.23	.39*	.37†	−.14	.35†	.43*	−.25	.36†	.42*	−.26	−.43*	−.28	−.06
Diff mean amp	D	.46*	.52**	.51**	−.35†	.69***	.66***	.68***	−.34†	.50**	.48*	.00	.49*	.55	−.10	.49*	.53**	−.11	−.36†	−.21	−.15
Diff AUC	D	.48*	.53**	.53**	−.33	.72***	.69***	.72***	−.29	.53**	.51**	.04	.51**	.58**	−.08	.52**	.57**	−.08	−.32	−.19	−.14
Diff peak lat	D	−.22	−.30	−.28	.48*	−.50*	−.57**	−.53**	.52**	−.16	−.14	−.04	−.18	−.19	.05	−.17	−.15	.05	.11	.44*	.19
Counting oddball																					
Peak amp	D	–	–	–	–	–	–	–	–	.70***	.66***	.47*	.70***	.70***	.42*	.71***	.69***	.43*	.11	−.02	.03
Mean amp	D	–	–	–	–	–	–	–	–	.68***	.63***	.45*	.66***	.67***	.42*	.67***	.65***	.42*	.02	−.12	−.01
AUC	D	–	–	–	–	–	–	–	–	.67***	.62***	.45*	.67***	.66***	.42*	.67***	.64***	.42*	.07	−.07	.03
Peak lat	D	–	–	–	–	–	–	–	–	−.22	−.23	−.39*	−.35†	−.27	−.41*	−.35†	−.26	−.40*	−.16	.09	.16
Diff peak amp	D	–	–	–	–	–	–	–	–	.65***	.64***	.32	.63**	.68**	.17	.64***	.67***	.18	−.04	−.18	−.01
Diff mean amp	D	–	–	–	–	–	–	–	–	.51**	.53**	.26	.53**	.62**	.15	.55**	.60**	.15	−.13	−.20	−.16
Diff AUC	D	–	–	–	–	–	–	–	–	.58**	.57**	.30	.60**	.66***	.21	.61**	.65***	.21	−.07	−.16	−.17
Diff peak lat	D	–	–	–	–	–	–	–	–	−.08	−.15	−.26	−.19	−.17	−.25	−.20	−.19	−.23	.04	.14	.38†

Note. AUC = area under the curve; Diff = difference; lat = latency; amp = amplitude; C = compatible; D = deviant; I = incompatible; N = No-Go. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

amplitude. Recording EEG during an SRC task also offers the possibility to assess LRPs. LRPs represent the preferential activation of one hemisphere relative to the other when a motor response is being prepared (Coles, 1989; Freeman et al., 2011; Roggeveen et al., 2007). To our knowledge, the test-retest reliability of LRP measures has never been studied. Our results showed that the LRP peak measure shows some level of test-retest reliability, with however large confidence intervals. Regarding onset latency, the compatible LRP onset showed poor test-retest reliability and thus only the incompatible onset was reliable. Since the compatible LRP onset occurs earlier than the incompatible onset, this finding might be explained by more intra-individual variability during earlier stages of motor processing. Given that this is the first investigation of the LRP test-retest reliability of LRPs and that our sample was rather small, these results must be interpreted with caution. Future research will be important to better pinpoint the reliability of LRPs, especially since they have a potential value as biomarkers of therapeutic improvement in clinical practice (Morand-Beaulieu et al., 2018).

Interestingly, strong correlations were found between components elicited by either version of the oddball task. This suggests that both tasks are sensitive to similar processes. Also, a smaller P3b amplitude was found in response to the counting oddball task, compared to the motor oddball task. Hill et al. (1995) also reported larger P3b amplitude in response to a motor than to a counting oddball task. However, our results are opposed to the findings of Salisbury et al. (2001). In their study, they argued that the motor variant of the oddball task required less resource allocation than silent counting of deviant stimuli. Yet, their participants were only asked to press a button for deviant stimuli, while our experiment required a motor response to standard and deviant stimuli. As previously pointed by Steinhauer and Hill (1993), task load differentially affects counting and motor oddball tasks. Therefore, it is plausible that our counting oddball task requires more cognitive resources such as working memory updating, compared to our motor variant.

Limitations and Future Perspectives

One of the main limitations of the current study is its small sample size. While it is comparable to other recent studies assessing the test-retest reliability of ERPs (Brunner et al., 2013; Cassidy et al., 2012; Groves et al., 2018; Weinberg & Hajcak, 2011), a larger sample would reduce the confidence intervals and yield more precise and robust reliability coefficients. Yet, ERP research is resource-demanding and it is hard for individual research groups to collect large samples. Therefore, meta-analyses of existing ERP data will allow a better understanding of test-retest reliability. The large age range of our sample constitutes another limitation. However, this should not have significantly affected our findings, since good test-retest reliability has been found across the lifespan (Hammerer et al., 2013; Walhovd & Fjell, 2002). A larger sample would also allow investigating the potential contribution of adult brain development on the test-retest reliability of various ERPs.

While larger samples could help to enhance reliability coefficients, other ERP scoring methods could possibly increase reliability, especially for ERPs with low-reliability coefficients. Methods such as principal (PCA) or independent component analysis (ICA) allow the data-driven identification of components without a priori assumptions and can distinguish spatially or temporally overlapping components (Beauducel & Debener, 2003; Makeig et al., 1997). For instance, PCA factor scores of the auditory P100, N100, and P200 were shown to be more temporally reliable than peak measures (Beauducel et al., 2000). However, a study reported that ICA-derived NoGo P300 components showed comparable test-retest reliability with the channel-derived NoGo P300 wave (Brunner et al., 2013). Thus, we could speculate that similar levels of test-retest reliability would be obtained had we used an ICA-based decomposition of ERP waveforms.

Conclusions

Our results showed that reliability coefficients were very similar in motor and counting oddball tasks. This suggests that motor responses do not significantly confound the test-retest reliability of the ERPs elicited during oddball tasks. In addition, ERPs elicited during the SRC paradigm have good test-retest reliability, especially for amplitude measures. These results confirm that ERPs have the potential to constitute robust markers of brain function and are well suited to serve as assessment tools in longitudinal or clinical studies.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/0269-8803/a000286

ESM 1. Supplementary results; mean number of included trials in each condition/component (Table E1)

ESM 2. Full correlation matrix of all variables in the study (Table E2)

References

American EEG Society. (1994). Guideline thirteen: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology, 11*(1), 111–113.

Beauducel, A., & Debener, S. (2003). Misallocation of variance in event-related potentials: Simulation studies on the effects of test power, topography, and baseline-to-peak versus principal component quantifications. *Journal of Neuroscience Methods, 124*(1), 103–112. https://doi.org/10.1016/S0165-0270(02)00381-3

Beauducel, A., Debener, S., Brocke, B., & Kayser, J. (2000). On the reliability of augmenting/reducing: Peak amplitudes and principal component analysis of auditory evoked potentials. *Journal of Psychophysiology, 14*(4), 226–240. https://doi.org/10.1027/0269-8803.14.4.226

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56*(6), 893–897.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Bledowski, C., Prvulovic, D., Hoechstetter, K., Scherg, M., Wibral, M., Goebel, R., & Linden, D. E. J. (2004). Localizing P300 generators in visual target and distractor processing: A combined event-related potential and functional magnetic resonance imaging study. *The Journal of Neuroscience, 24*(42), 9353–9360. https://doi.org/10.1523/JNEUROSCI.1897-04.2004

Brunner, J. F., Hansen, T. I., Olsen, A., Skandsen, T., Håberg, A., & Kropotov, J. (2013). Long-term test-retest reliability of the P3 NoGo wave and two independent components decomposed from the P3 NoGo wave in a visual Go/NoGo task. *International Journal of Psychophysiology, 89*(1), 106–114. https://doi.org/10.1016/j.ijpsycho.2013.06.005

Cassidy, S. M., Robertson, I. H., & O'Connell, R. G. (2012). Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology, 49*(5), 659–664. https://doi.org/10.1111/j.1469-8986.2011.01349.x

Coles, M. G. (1989). Modern mind-brain reading: Psychophysiology, physiology, and cognition. *Psychophysiology, 26*(3), 251–269.

Conners, C. K., Sitarenios, G., & Ayearst, L. E. (2018). Conners' Continuous Performance Test. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (3rd ed., pp. 929–933). Springer International Publishing.

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*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(4), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Dunn, O. J., & Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association, 66*(336), 904–908. https://doi.org/10.1080/01621459.1971.10482369

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Brooks/Cole Publishing.

Fallgatter, A. J., Aranda, D. R., Bartsch, A. J., & Herrmann, M. J. (2002). Long-term reliability of electrophysiologic response control parameters. *Journal of Clinical Neurophysiology, 19*(1), 61–66.

Fallgatter, A. J., Brandeis, D., & Strik, W. K. (1997). A robust assessment of the NoGo-anteriorisation of P300 microstates in a cued Continuous Performance Test. *Brain Topography, 9*(4), 295–302. https://doi.org/10.1007/bf01464484

Fallgatter, A. J., & Strik, W. K. (1999). The NoGo-anteriorization as a neurophysiological standard-index for cognitive response control. *International Journal of Psychophysiology, 32*(3), 233–238. https://doi.org/10.1016/S0167-8760(99)00018-5

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32. https://doi.org/citeulike-article-id:2346712

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45*(1), 152–170. https://doi.org/10.1111/j.1469-8986.2007.00602.x

Ford, J. M., Mathalon, D. H., White, P. M., & Pfefferbaum, A. (2000). Left temporal deficit of P300 in patients with schizophrenia: Effects of task. *International Journal of Psychophysiology, 38*(1), 71–79. https://doi.org/10.1016/S0167-8760(00)00131-8

Freeman, J. B., Ambady, N., Midgley, K. J., & Holcomb, P. J. (2011). The real-time link between person perception and action: Brain potential evidence for dynamic continuity. *Social Neuroscience, 6*(2), 139–155. https://doi.org/10.1080/17470919.2010.490674

Fruehwirt, W., Dorffner, G., Robert, S., Gerstgrasser, M., Grossegger, D., Schmidt, R., Dal-Bianco, P., Ransmayr, G., Garn, H., Waser, M., & Benke, T. (2018). Associations of event-related brain potentials and Alzheimer's disease severity: A longitudinal study. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 92*, 31–38. https://doi.org/10.1016/j.pnpbp.2018.12.013

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*(4), 468–484.

Groves, K., Kennett, S., & Gillmeister, H. (2018). Early visual ERPs show stable body-sensitive patterns over a 4-week test period. *PLoS One, 13*(2), Article e0192583. https://doi.org/10.1371/journal.pone.0192583

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

Hammerer, D., Li, S. C., Volkle, M., Muller, V., & Lindenberger, U. (2013). A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. *Psychophysiology, 50*(1), 111–123. https://doi.org/10.1111/j.1469-8986.2012.01476.x

Herrmann, C. S., & Knight, R. T. (2001). Mechanisms of human attention: Event-related potentials and oscillations. *Neuroscience & Biobehavioral Reviews, 25*(6), 465–476.

Hill, S. Y., Steinhauer, S., & Locke, J. (1995). Event-related potentials in alcoholic men, their high-risk male relatives, and low-risk male controls. *Alcoholism: Clinical and Experimental Research, 19*(3), 567–576.

Hittner, J. B., May, K., & Silver, N. C. (2004). Testing dependent correlations with nonoverlapping variables: A Monte Carlo simulation. *The Journal of Experimental Education, 73*(1), 53–69. https://doi.org/10.3200/JEXE.71.1.53-70

Houston, R. J., & Schlienz, N. J. (2018). Event-related potentials as biomarkers of behavior change mechanisms in substance use disorder treatment. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(1), 30–40. https://doi.org/10.1016/j.bpsc.2017.09.006

Huster, R. J., Westerhausen, R., Pantev, C., & Konrad, C. (2010). The role of the cingulate cortex as neural generator of the N200 and P300 in a tactile response inhibition task. *Human Brain Mapping, 31*(8), 1260–1271. https://doi.org/10.1002/hbm.20933

Ishihara, S. (1917). *Tests for color-blindness*. Hongo Harukicho.

Kayser, J., Tenke, C. E., Gil, R., & Bruder, G. E. (2010). ERP Generator patterns in schizophrenia during tonal and phonetic oddball tasks: Effects of response hand and silent count. *Clinical EEG and Neuroscience, 41*(4), 184–195.

Kinoshita, S., Inoue, M., Maeda, H., Nakamura, J., & Morita, K. (1996). Long-term patterns of change in ERPs across repeated measurements. *Physiology & Behavior, 60*(4), 1087–1092. https://doi.org/10.1016/0031-9384(96)00130-8

Kok, A. (1988). Overlap between P300 and movement-related-potentials: A response to Verleger. *Biological Psychology, 27*(1), 51–58.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kutas, M., Kiang, M., & Sweeney, K. (2012). Potentials and paradigms: Event-related brain potentials and neuropsychology. In D. I. Mostofsky & M. Faust (Eds.), *The handbook of the neuropsychology of language* (pp. 543–564). Wiley-Blackwell.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. MIT Press.

Maeda, H., Morita, K., Nakamura, J., Inoue, M., Kinoshita, S., Kodama, E., Maki, S., & Nakazawa, Y. (1995). Reliability of the task-related component (P3b) of P3 event-related potentials. *Psychiatry and Clinical Neurosciences, 49*(5–6), 281–286.

Makeig, S., Jung, T. P., Bell, A. J., Ghahremani, D., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences of the USA, 94*(20), 10979–10984.

Martínez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., Wong, E. C., Hinrichs, H., Heinze, H. J., & Hillyard, S. A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience, 2*, 364–369. https://doi.org/10.1038/7274

Mathalon, D. H., Ford, J. M., & Pfefferbaum, A. (2000). Trait and state aspects of P300 amplitude reduction in schizophrenia: A retrospective longitudinal study. *Biological Psychiatry, 47*(5), 434–449. https://doi.org/10.1016/S0006-3223(99)00277-2

Morand-Beaulieu, S., O'Connor, K. P., Blanchet, P. J., & Lavoie, M. E. (2018). Electrophysiological predictors of cognitive-behavioral therapy outcome in Tic disorders. *Journal of Psychiatric Research, 105*, 113–122. https://doi.org/10.1016/j.jpsychires.2018.08.020

Morand-Beaulieu, S., O'Connor, K. P., Richard, M., Sauve, G., Leclerc, J. B., Blanchet, P. J., & Lavoie, M. E. (2016). The impact of a cognitive-behavioral therapy on event-related potentials in patients with Tic disorders or body-focused repetitive behaviors. *Frontiers in Psychiatry, 7*, Article 81. https://doi.org/10.3389/fpsyt.2016.00081

Morgan, K. K., Luu, P., & Tucker, D. M. (2016). Changes in P3b latency and amplitude reflect expertise acquisition in a football visuomotor learning task. *PLoS One, 11*(4), Article e0154021. https://doi.org/10.1371/journal.pone.0154021

Munsters, N. M., van Ravenswaaij, H., van den Boomen, C., & Kemner, C. (2019). Test-retest reliability of infant event related potentials evoked by faces. *Neuropsychologia, 126*, 20–26. https://doi.org/10.1016/j.neuropsychologia.2017.03.030

Parvaz, M. A., Maloney, T., Moeller, S. J., Malaker, P., Konova, A. B., Alia-Klein, N., & Goldstein, R. Z. (2014). Multimodal evidence of regional midcingulate gray matter volume underlying conflict monitoring. *NeuroImage: Clinical, 5*, 10–18. https://doi.org/10.1016/j.nicl.2014.05.011

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. H. K. Lewis.

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. https://doi.org/10.1016/j.biopsycho.2013. 04.007

Roggeveen, A. B., Prime, D. J., & Ward, L. M. (2007). Lateralized readiness potentials reveal motor slowing in the aging brain. *The Journals of Gerontology: Series B, 62*(2), P78–P84. https://doi.org/10.1093/geronb/62.2.P78

Salisbury, D. F., Rutherford, B., Shenton, M. E., & McCarley, R. W. (2001). Button-pressing affects P300 amplitude and scalp topography. *Clinical Neurophysiology, 112*(9), 1676–1684.

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*(8), 1427–1437. https://doi.org/10.1016/S1388-2457(00)00320-5

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology, 30*(5), 451–459.

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*(2), 260–270. https://doi.org/10.1111/j.1469-8986.2009.00942.x

Simon, J. R., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics, 6*(1), 99–105. https://doi.org/10.1080/00140136308930679

Smulders, F. T., Kenemans, J. L., & Kok, A. (1996). Effects of task variables on measures of the mean onset latency of LRP depend on the scoring method. *Psychophysiology, 33*(2), 194–205.

Snellen, H. (1862). *Probebuchstaben zur Bestimmung der Sehschärfe* [Sample letters to determine the visual acuity]. P. W. van de Weijer.

Steinhauer, S. R., & Hill, S. Y. (1993). Auditory event-related potentials in children at high risk for alcoholism. *Journal of Studies on Alcohol and Drugs, 54*(4), 408–421.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS One, 8*(9), Article e73990. https://doi.org/10.1371/journal.pone.0073990

Wachinger, C., Volkmer, S., Bublath, K., Bruder, J., Bartling, J., & Schulte-Körne, G. (2018). Does the late positive component reflect successful reading acquisition? A longitudinal ERP study. *NeuroImage: Clinical, 17*, 232–240. https://doi.org/10.1016/j.nicl.2017.10.014

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*(1), 29–40. https://doi.org/10.1016/S0167-8760(02)00039-9

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

Williams, L. M., Simms, E., Clark, C. R., Paul, R. H., Rowe, D., & Gordon, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "Neuromarker". *International Journal of Neuroscience, 115*(12), 1605–1630. https://doi.org/10.1080/00207450590958475

Wronka, E., Kaiser, J., & Coenen, A. M. (2008). The auditory P3 from passive and active three-stimulus oddball paradigm. *Acta Neurobiologiae Experimentalis (Warsaw), 68*(3), 362–372.

History
Received January 12, 2021
Revision received May 11, 2021
Accepted June 18, 2021
Published online August 25, 2021

Acknowledgments
We want to thank all participants who took part in the current study. We also wish to express our gratitude to Martine Germain for her assistance with electrophysiological recordings.

Conflict of Interest
The authors report no conflict of interest related to this work.

Publication Ethics
This study was approved by the ethics committee of the Institut Universitaire en Santé Mentale de Montréal (#2012-029) and informed consent was obtained from all participants prior to their participation in the study.

Funding
This project was funded by a grant from the Canadian Institutes for Health Research (CIHR; #93556) awarded to Marc E. Lavoie. Simon Morand-Beaulieu was supported by a doctoral scholarship from the FRQS (#32114), the Robert-Élie doctoral award from the Centre de recherche de l'Institut Universitaire en Santé Mentale de Montréal, and a postdoctoral fellowship award from the Canadian Institutes of Health Research (#415541).

Simon Morand-Beaulieu
Child Study Center
Yale University School of Medicine
230 South Frontage Road
New Haven, CT 06520
USA
simon.morand-beaulieu@yale.edu