IOP Publishing J. Neural Eng. 17 (2020) 036024 https://doi.org/10.1088/1741-2552/ab85d4

Journal of Neural Engineering

RECEIVED
4 October 2019

REVISED
25 March 2020

ACCEPTED FOR PUBLICATION
2 April 2020

PUBLISHED
12 June 2020

PAPER

Computer data simulator to assess the accuracy of estimates of visual N2/N2pc event-related potential components

Francesca Marturano[1], Sabrina Brigadoi[1,2], Mattia Doro[2], Roberto Dell'Acqua[2,3] and Giovanni Sparacino[1,4]

[1] Department of Information Engineering—DEI, University of Padova, Padova, Italy
[2] Department of Developmental Psychology—DPSS, University of Padova, Padova, Italy
[3] Padova Neuroscience Center, University of Padova, Padova, Italy
[4] Author to whom any correspondence should be addressed.

E-mail: gianni@dei.unipd.it

Keywords: visual N2/N2pc, ERP, conventional average, Gaussian mixture models, simulation

Abstract

Objective. Event-related potentials (ERPs) evoked by visual stimulations comprise several components, with different amplitudes and latencies. Among them, the N2 and N2pc components have been demonstrated to be a measure of subjects' allocation of visual attention to possible targets and to be involved in the suppression of irrelevant items. Unfortunately, the N2 and N2pc components have smaller amplitudes compared with those of the background electroencephalogram (EEG), and their measurement requires employing techniques such as conventional averaging, which in turn necessitates several sweeps to provide acceptable estimates. In visual search studies, the number of sweeps (*Nswp*) used to extrapolate reliable estimates of N2/N2pc components has always been somehow arbitrary, with studies using 50–500 sweeps. *In-silico* studies relying on synthetic data providing a close-to-realistic fit to the variability of the visual N2 component and background EEG signals are therefore needed to go beyond arbitrary choices in this context. *Approach.* In the present work, we sought to take a step in this direction by developing a simulator of ERP variations in the N2 time range based on real experimental data while monitoring variations in the estimation accuracy of N2/N2pc components as a function of two factors, i.e. signal-to-noise ratio (SNR) and number of averaged sweeps. *Main results.* The results revealed that both *Nswp* and SNR had a strong impact on the accuracy of N2/N2pc estimates. Critically, the present simulation showed that, for a given level of SNR, a non-arbitrary *Nswp* could be parametrically determined, after which no additional significant improvements in noise suppression and N2/N2pc accuracy estimation were observed. *Significance.* The present simulator is thought to provide investigators with quantitative guidelines for designing experimental protocols aimed at improving the detection accuracy of N2/N2pc components. The parameters of the simulator can be tuned, adapted, or integrated to fit other ERP modulations.

1. Introduction

Event-related potentials (ERPs) provide important information on brain functioning, in both normal and pathological conditions [1–4]. Therefore, their correct estimation is fundamental in cognitive studies and in studies aiming to develop brain-computer interface (BCI) applications, which can translate brain activity into computer commands [5–7]. The accuracy of ERP estimation is inevitably bound to its small amplitude in relation to the background electroencephalographic (EEG) noise, which makes the estimation of an ERP from EEG activity recorded on a single trial almost impossible. In fact, ERPs are commonly estimated by averaging EEG activity recorded on tens if not hundreds of trials, so as to bring to a minimum the influence of EEG noise on ERP isolation [8, 9]. The definition of an optimal estimation method goes hand-in-hand with the need for high reliability. Factors such as the experimental setup, user proficiency in collecting the data, and noise sources could affect the reliability of the ERP estimate.

Many studies in the ERP field have attempted to validate novel ERP estimation techniques by either

developing ad hoc simulation frameworks or using existing tools to generate and/or assess EEG data [10–12]. Simulation creates an environment that is fully under the user's control, so that several factors can be manipulated in a controlled fashion and their contribution to the reliability of an ERP estimate precisely evaluated. Crucial aspects of the design of a simulation framework include, firstly the capability of reliably mimicking the shape and parameters of the ERP (e.g. number and shape of the components, latencies, and amplitudes) and the effects of the conditions that may occur in real practice (e.g. variability, adaptation, and noise content). An additional value of a simulation framework is its flexibility, i.e. the possibility of being adapted/extended by other researchers to deal with problems belonging to a class broader than the one originally considered. Kiesel and colleagues [13], for example, created a simulation directly from real data by introducing a known effect on the latency of a subset of ERP components (i.e. N1, P3, N2pc, and P3b) to validate the performance of different techniques estimating the latency onset. Although Kiesel *et al's* simulation approach preserved the characteristics of a real context, it did not implement explicit solutions to control for factors that commonly affect an ERP estimate, like EEG background noise. Other simulation techniques, on the other hand, have developed specific ERP modeling frameworks [14, 15], which lacked a realistic reproduction of the inter-trial ERP variability that is commonly observed in every experiment or did not take into consideration factors that could corrupt the measurements, like the level of background EEG noise. These aspects are, instead, fundamental when simulation is used to validate ERP estimation methods. More sophisticated, multi-components, and integrated tools to generate and process the EEG have also been proposed. The rationale underlying these approaches is to reproduce the inner brain sources and the neural connections that cooperate to generate the evoked ERP. The definition of this cooperation depends on the specific process and approach adopted, and requires studies of connectivity and/or source localization, behavioral analysis, head models, and an accurate physiological and psychological knowledge of the brain processes involved. The final model is composed by several interacting sub-models, each of which simulates a peculiar component of the chain of mechanisms yielding the evoked response [16–19]. For example, Tan and Wyble [19] developed a model of the neural mechanisms that generate the N2pc component during a target localization process. The model was a replication of the visual system and was composed of three interconnected layers of neurons with a receptive field getting larger between earlier and later layers. The first layer represented the early visual area receiving the input from the environment, the second the late visual area, and the third the attention map that corresponded to the brain areas responsible for visual attention deployment. The activity of each layer excited a localized group of neurons in the following layer until reaching the attention map. The amplitude of the ERP was generated at any given time point in proportion to the number of activated neurons in an underlying cerebral region. In the model, when a neuron's membrane potential is excited above a threshold, the simulated EEG signal shifts away from baseline, whereas the inhibition of the neuron causes the EEG to return to baseline. In their work, the development of this model is preceded by an accurate psychological study demonstrating that N2pc reflects neural processes first involved in spatially localizing a target and then deploying attention to it, rather than a concomitant interaction between target enhancement and distractor suppression. Lindgren *et al* [12], instead, implemented a different approach and created a framework, named *sim*BCI, that integrates several components, including a head model, a source generator, and a model of the brain processes, aimed at creating arbitrary BCI experiments where event timelines, artefacts, and other parameters can be set by users from a high level interface. Although these tools appear to be powerful, it seems extremely long and complex to generalize them to new cognitive processes or components, since a deep physiological and psychological knowledge of the process to model is required and not always available.

To overcome the limitations and drawbacks of the previously presented approaches, in this work we developed a simulator of ERP signals emulating the variability present in real experimental data. We focused on describing the typical modulations of the N2 and N2pc ERP components elicited in visual search tasks. In these tasks, participants are required to search for a laterally displayed target object embedded among distractors while gazing at a central fixation point aligned to the sagittal axis. The successful selection of the target produces modulations in the N2 parameters [20–23]. The ERP response recorded at parieto-occipital sites (e.g. PO7/8) usually differs between electrodes located contralaterally (CL) and ipsilaterally (IL) to the lateral target in the N2 time window, i.e. 200–300 ms (figure 1). Specifically, CL activity is more negative than IL activity. Consequently, the amplitude of the N2pc, computed by subtracting the IL ERP response from the CL ERP response [20, 24], increases indicating the subject's allocation of attention to the lateral target. For this reason, N2pc has been widely employed to study the mechanisms that guide visual attention in space [25–27] and in recent years several studies have started exploring its use in BCI applications [5, 7, 28].

However, N2 and N2pc are characterized by a very small amplitude (usually smaller than 2 μV), compared with the P3 component, for example. Therefore, their estimation is more challenging due to the lower signal-to-noise ratio (SNR) of the measured signal. A large number of sweeps (*Nswp*) should

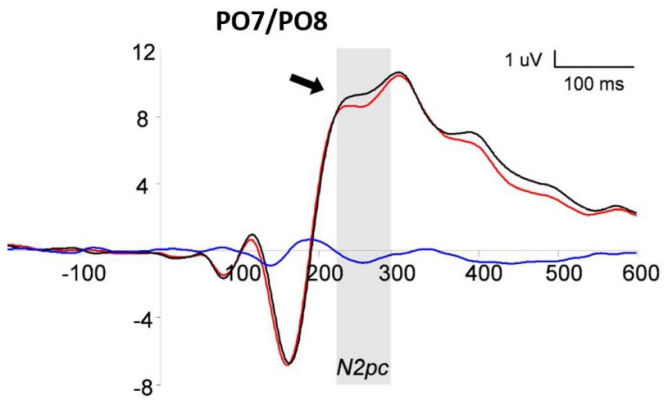

Figure 1. The N2 component on the contralateral to the target location (red line) is compared to the N2 component on the ipsilateral location (black line). The N2pc (blue line) is obtained by subtracting the ipsilateral signal from the contralateral one.

(generally) be averaged together to have enough SNR to estimate reliable effects. In the literature, this value is however extremely variable, ranging from approximately 60 sweeps [9] to 512 sweeps [29], according to the personal experience of the investigators or to values used in previous studies. The *a-priori* selection of the most suitable *Nswp* to provide a reliable estimate is not a simple task because the ideal *Nswp* can depend on several factors, such as the noise of a given experimental setting, the nature and amplitude of the ERP of interest, the expected difference in amplitude between conditions, the experiment duration, and/or the participant's fatigue [30]. Previous studies have tried to tackle this important aspect and put forth approaches for establishing the minimum *Nswp* to obtain a stable and reliable estimate of the ERP component of interest. For instance, in some studies investigating components like the error-related negativity, the late positive potential, or the P300 component, the *Nswp* was determined based on the logic of producing a 'high correlation' between an ERP obtained with relatively few sweeps and an ERP obtained using more sweeps [31–34]. In a recent study, Boudewyn and colleagues [35] emphasized that the ideal *Nswp* should not overlook the statistical power of the estimate, suggesting that it should be decided after a careful analysis of the noise content of the data, the sample size, and the amplitude of the effects other than the stability of the ERP estimate.

In this work, the accuracy of the estimate of the ERP was evaluated by using a reliable and reproducible simulation method, where experimental protocol variables such as the power of the background EEG noise, the *Nswp*, and the magnitude of the ERP were separately modulated. We generated synthetic, though realistic, ERP/EEG datasets, based on real experimental data collected from human volunteers, and used simulation to assess the reliability of the estimation of the N2 and N2pc components. The purpose of the present study is essentially two-fold: (1) to propose a novel and simple computer simulation framework capable of realistically mimicking the N2/N2pc components and (2) to use synthetic data to evaluate the accuracy of N2 and N2pc estimation employing the most used ERP estimation method, i.e. conventional averaging (CA), when varying the *Nswp* and SNR of the data. As far as the first aim is concerned, the main novelty is that we simulated the data considering the variability of the N2 component as estimated on real data and corrupting the recordings using real background EEG additive noise measured in a resting state (RS). As far as the second aim is concerned, we investigated three synthetic datasets with different SNRs (obtained by perturbing in a controlled way the statistical features of the RS noise added to each ERP) and evaluated the accuracy of CA depending on the *Nswp* and the level of SNR. The proposed approach could be easily reproduced, optimized, and extended to protocols designed to explore ERP components, thus offering a level of flexibility higher than previously published approaches.

2. Methods

2.1. Design of the proposed simulator

Figure 2 summarizes the steps taken to develop the present synthetic scenario. The block diagram in figure 2 will be used throughout the forthcoming sections as a roadmap to describe the various passages for the generation of the present synthetic framework, each of which will be referenced in the corresponding section using the box labels 'A' to 'I' for brevity. To simulate a realistic EEG dataset acquired during a visual search task, we deemed of critical importance to first evaluate the real variability of the elicited N2 component across participants and estimate a common template for the ERP. To this aim, the variability of the N2/N2pc components was estimated using a real EEG dataset collected from participants performing in a visual search task designed to elicit an N2/N2pc response.

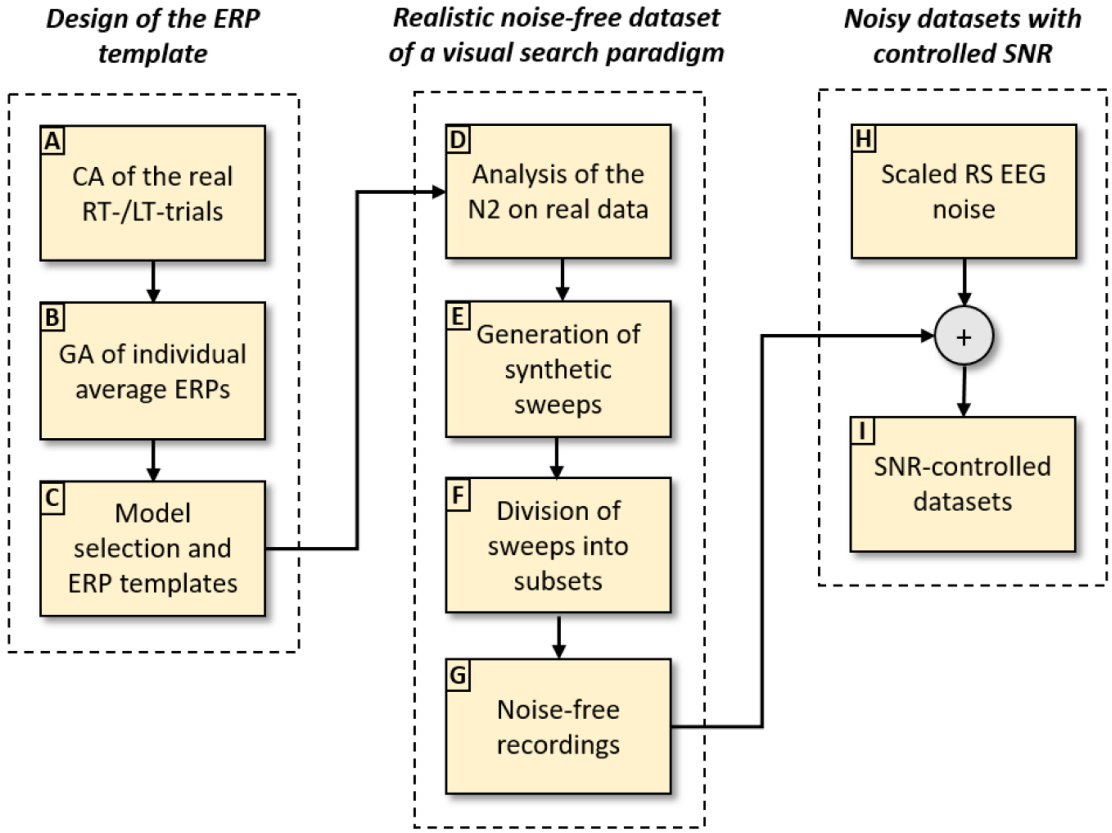

Figure 2. Block diagram of the proposed approach to simulate a realistic visual search dataset eliciting lateralized modulations of the N2 ERP component under variable SNR conditions. CA = conventional average; GA = group average; RS = resting state. Blocks 'A' to 'I' are referenced in the text to illustrate the corresponding mathematical routine.

We then created a synthetic noise-free dataset that emulated the real variability of the N2 component. An RS EEG noise was then added to the noise-free dataset controlling the range of the resulting SNR, by modulating the noise amplitude with multiplicative factors. By doing so, we created three synthetic noisy scenarios, with different SNRs, to simulate possible experimental conditions achieved with different paradigms and in different laboratories.

2.1.1. Participants and procedure.
Fourteen healthy participants (mean age 23.2 ± 2.1 years, 5 males) took part in an experiment employing a common visual search paradigm eliciting lateralized modulations of the N2 component due to attention deployment to lateral targets embedded among distractors. All subjects had normal or corrected-to-normal vision. The experiment lasted about 40 min and was divided into several blocks. Each trial was started with a spacebar press, so as to allow participants to take short breaks throughout the experiment to attenuate effects induced by physical and/or mental fatigue. The study was approved by the ethical committee of the University of Padova.

An example of the stimuli and a schematic illustration of the sequence of events on each trial of the visual search paradigm are illustrated in figure 3.

Each trial began with the presentation of a centrally displayed fixation cross for 750–950 ms (randomly jittered in 50 ms steps), followed by the presentation of the visual search array for 150 ms. Each visual array was composed of 12 white circles regularly spaced around a notional clock face with a ray of 2.5° of visual angle. The visual array was displayed against the black background of a 25″ CRT computer monitor with a 60 Hz refresh rate, at a viewing distance of approximately 60 cm. A white bar was presented inside each circle, which could be tilted or vertical with equal probability. One of the circles—the target circle—was blue, and was positioned in one-third of trials either above or below fixation (i.e. aligned to the sagittal midline), or in one of the lateral positions in the remaining two-thirds of the trials, with equal probability to the left or right of central fixation. The participants had to report the orientation of the bar within the target circle by pressing one of two keys on the computer keyboard (i.e. '1' or '2'), using the index or middle finger of their right hand, respectively. The maximum time for responding was 1500 ms. The participants were instructed to keep their gaze centrally fixated throughout each trial and to respond as fast and accurately as possible. Each participant ran through 600 trials, namely, 400 trials with lateral—left or right—targets, and 200 trials with midline targets.

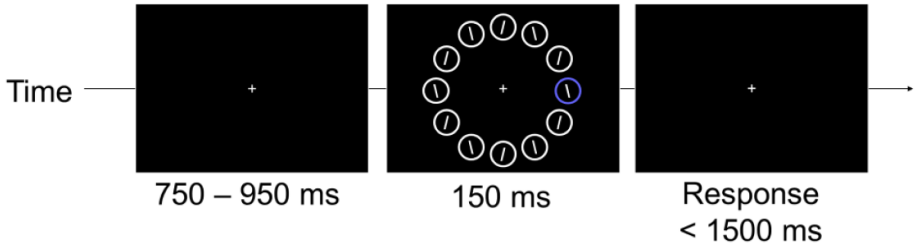

Figure 3. Design of the experimental paradigm to elicit the N2. Each trial began with the presentation of a centrally displayed fixation cross for 750–950 ms (left panel), followed by the presentation of the visual search array for 150 ms (central panel). The array was composed of 12 white circles regularly spaced around a notional clock face. A white bar, either tilted or vertical, was presented inside each circle. One of the circles, the target one, was instead blue. After the visual search array disappeared, the participants had 1500 ms to report the orientation of the bar inside the target circle, while viewing a black screen with the central fixation cross (right panel).

2.1.2. Data acquisition.
The EEG data was acquired from 28 scalp electrodes positioned on an elastic cap according to the international 10–20 system. Three additional electrodes, placed at the outer canthi and below the left eye, were used to register horizontal (HEOG) and vertical (VEOG) eye movements. The EEG activity was amplified and digitized at a sampling rate of 500 Hz, referenced to the activity recorded on the left earlobe, and then re-referenced offline to the average of the left and right earlobes electrodes [36, 37].

2.1.3. EEG data pre-processing.
The EEG data were analyzed with a standard pre-processing pipeline using 'EEGLAB' (version 14.1.2, MATLAB version 2017b, The MathWorks, Natick, 2017) [38]. The data were band-pass filtered (filter bandwidth of 0.1–30 Hz), and only the signals from the channels of interest (i.e. P7 and P8) were selected for the subsequent analyses [25, 39–42]. The continuous EEG was segmented in sweeps starting at 200 ms before the visual array onset and ending 1000 ms after to largely include the typical time window of the N2/N2pc components, i.e. about 200–300 ms [20, 27]. Epochs were baseline corrected using the average activity in the time interval starting from −200 ms and the visual array onset. The VEOG channel was computed as the difference between Fp1 and the electrode below the left eye, since blink episodes more strongly affect the frontal sites [41, 43, 44]. The HEOG channel was computed as the subtraction of the two electrodes placed at the outer canthi. Sweeps with artefacts (eye blinks exceeding 20 μV in the VEOG channel, saccades exceeding 50 μV in the HEOG channel, or muscular artefacts exceeding 60 μV in all other channels) were excluded. Sweeps associated with incorrect responses were discarded, as well as sweeps associated with midline target sweeps. The remaining sweeps were divided according to the target side, i.e. into right-target (RT) and left-target (LT) sweeps, to divide the contribution of CL and IL scalp sites (e.g. for RT sweeps, the elicited CL activity is measured at P7, whereas the IL is measured at P8). This subdivision created four groups of sweeps, hereafter labeled as P7R, P7L, P8R, and P8L. In this example, the label P7R refers to the sweeps measured at P7 when right targets were presented. A mean of about 100 sweeps was available for each participant and condition, ranging from a minimum of 23 sweeps (subject n°1, RT condition) and a maximum of 176 sweeps (subject n°3, LT condition). The averaged ERPs were computed for each group of sweeps and participant (figure 2, block A).

2.1.4. Design of the ERP template.
For each group, a grand-averaged (GA) ERP was obtained by averaging the individual mean ERP values across participants (figure 2, block B). The GA ERP was used to build the ERP template. A Gaussian mixture model (GMM) was used to fit the GA ERP of each group (figure 2, block C), as already employed for this purpose [14, 15, 45, 46]. The flexibility of this model allows for a suitable fit of the peaks and valleys typical of ERP morphology [47, 48]. The only user-selected parameter of the model is the number of Gaussians. As the visual cognitive ERP is expected to be composed of at least five components (i.e. the positive P1, P2, and P3 and the negative N1 and N2) [47], a minimum of 5 Gaussians should be used in the model to fit these oscillations. Therefore, the GMM was fitted to the GA ERP using either 6, 7, or 8 Gaussians. The best-fitting model was selected based on the Akaike information criterion [49]. Data fitting was conducted using a literature-guided prior analysis of the expected position of the ERP voltage deflections [48]. The MATLAB *lsqnonlin* function was used to implement a non-linear least squares estimation of the model parameters. The best-fitting model was the 7-GMM, where 5 Gaussians captured the typical main ERP components, as expected from prior studies, whereas the last two fitted the tail of the ERP (figure 4(a)). Four different models (hereafter named templates) were obtained after fitting the RT/LT ERP of P7 and P8 (see figure 4(b)).

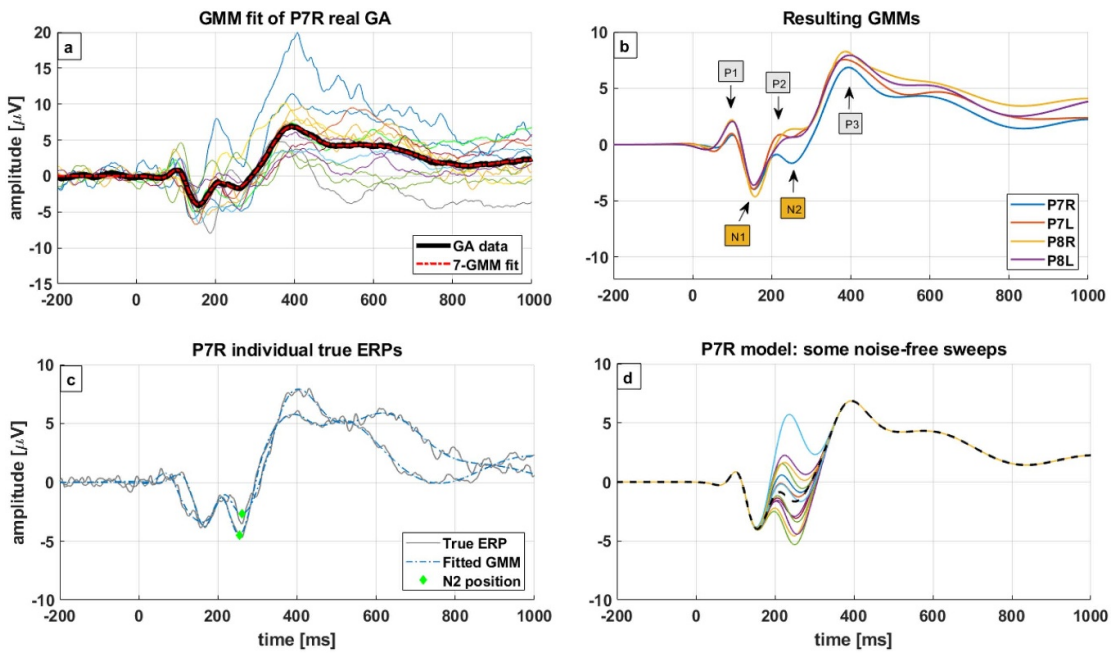

Figure 4. Illustrative steps to build realistic synthetic sweeps with variability in the N2 range. (a) Colored lines represent the individual average ERPs of the 14 subjects; the solid black line indicates the GA, while the dash-dotted red line is the fitted 7-Gaussian mixture model (GMM). These exemplary ERPs refer to the P7R group. (b) Resulting GMMs of the four groups. Colored boxes indicate the main expected components of the visual ERP, i.e. P1, P2, P3, N1, and N2. (c) Examples of how the N2 peak was identified from averaged ERPs of two representative participants. (d) Example of a few sweeps generated by perturbing the GMM derived from P7R data to generate a controlled variability of the P2/N2 component.

2.1.5. Realistic noise-free dataset of a visual search paradigm.

The four templates were first employed as an aid to estimate the variability of the N2 component in the real dataset, and then as models to be perturbed (in the N2 time range) to create the synthetic sweeps. Individual average ERPs of each group were fitted with the corresponding 7-GMM model to obtain a cleaner and more reliable estimate of the N2 peak amplitude and latency (figure 4(c) depicts two examples). For each fitted individual model, the maximum and minimum peak and the latency in the 200–300 ms time window were estimated (corresponding to the P2 and N2 components of the visual ERP; figure 2, block D).

Subsequently, the four templates were employed to create the synthetic noiseless sweeps. To generate many single sweeps with intra-individual variability in the N2 time range, two further and randomly perturbed Gaussians were added in the 200–300 ms time range of each template. The first Gaussian peak was positive and generated variability of amplitude and latency of the P2 component, while the second peak was negative and generated variability of amplitude and latency of the N2 component. Each template was used to generate more than 500 different sweeps (figure 2, block E). The variability of the N2 component was controlled so as to fit with the variability in the real dataset (figure 4(d)).

To recreate a standard dataset measured during a visual search experiment (i.e. both P7 and P8 signals), the synthetic single sweeps generated from each template were arranged to form three subsets differing in the N2 relative amplitude between CL and IL signals, hereafter labelled as S1, S2, and S3 (figure 2, block F). As a reminder, CL sweeps are those recorded at P7 for RT trials and those recorded at P8 for LT trials, whereas IL sweeps are the opposite. The sweeps of S1 matched the expected standard situation of $N2_{CL} < N2_{IL}$, whereas those of S2 and S3 reproduced unexpected yet possible situations, namely $N2_{CL} > N2_{IL}$ and $N2_{CL} \approx N2_{IL}$, respectively. A further complexity of the simulated dataset was the requirement of maintaining an internal consistency not only within each channel concerning the N2 amplitude for CL and IL sweeps (e.g. for S1, within P7, RT sweeps should have a more negative N2 than LT sweeps), but also a between-channel consistency (e.g. for S1, the channel CL to the target visual hemifield should present a more negative N2 relative to the IL channel). These subsets were organized to contain a pool of random sweeps that respected the imposed conditions on the CL and IL N2 amplitudes, both between and within channels. The random merging of these subsets gave rise to the noiseless synthetic dataset (figure 2, block G). At this stage, the number of simulated sweeps, as well as the percentage of sweeps randomly extracted from these subsets were user selected. In this illustrative study, we simulated 14 participants. For each of them, we created a synthetic noise-free recording of 200 sweeps, equally divided into RT and LT sweeps, for both channel P7 and P8. For each participant, 70

± 3% of the sweeps were randomly selected from S1, 15 ± 3% from S2, and 15 ± 3% from S3.

2.1.6. Noisy datasets with a controlled SNR.
The spontaneous EEG noise was generated based on RS recordings of two participants whose data were not part of the initial dataset. These data were acquired using the same experimental setup described in section 2.1.1, while participants were staring at a black screen for 5 min. The same pre-processing pipeline described in section 2.1.3 was applied to the RS recordings for all the 28 available scalp channels. The pre-processed signals were segmented to create a noise matrix containing sweeps lasting 1.2 s, the same length as the simulated sweeps.

To introduce variability in the noise power, the amplitude of all the RS sweeps was modulated by multiplying the matrix by five scale factors, i.e. 0.25, 0.5, 0.75, 1, and 1.25, thus obtaining five different noise matrices (figure 2, block H) [14, 50]. All the resulting noise sweeps were then sorted by increasing order of power. Noise sweeps with power lower than 1 μV^2 or higher than 300 μV^2 were discarded, as they were considered either unrealistic, or corrupted by long drifts or eye blinks [51].

Afterwards, the obtained sweeps of spontaneous EEG noise were summed to the noiseless sweeps of each participant, assuming an additive measurement model [14, 15, 45, 50, 52]. Three datasets were created from the same noise-free recordings of the 14 simulated subjects, by adding the EEG noise at different SNRs in each dataset (figure 2, block I). The SNR of the *i*th sweep was computed in a time window including the N2 component (i.e. 190–330 ms), as the ratio between the power of the noise-free ERP (u_i) and the power of the background noise (v_i) [14, 45, 53], as shown in equation (1):

$$SNR_i = \frac{P_{N2}(u_i)}{P_{N2}(v_i)}, \quad for\ i = 1:100, \tag{1}$$

where P_{N2} indicates the power of the signal, computed as the mean of squares of the data points in the N2 time window.

The noise sweeps were randomly and iteratively selected from the generated noise matrix, while controlling the SNR of the resulting sweep to be constantly within the range selected for that dataset. In this study, the three selected SNR ranges were: [0–0.4], [0.4–0.8], and [0.8–1.2]. For each simulated dataset, the signals were then pre-processed using the same pre-processing pipeline used for real data (see section 2.1.3).

The choice of the three SNR ranges was driven by the evaluation of the SNR distribution of the N2 component on the real dataset, using equation (1). In particular, the power of the useful signal in the numerator was estimated from the model fitted on the individual average ERP considering the N2 time range, whereas the power of each non-modulated RS sweep computed in the same time range was considered for the denominator, to obtain a raw distribution of the possible SNR values on real data. The estimated SNR distribution in the N2 time window resulted in a median of 0.18 (0.06–0.5, 25th–75th percentiles, respectively), with the lowest and highest adjacent values of 8×10^{-4} and 1.12, respectively.

2.2. Use of the simulator to assess the accuracy of average N2/N2pc components estimation

We used the three SNR-controlled datasets to assess the performance of the most basic and popular method employed to estimate the ERPs related to the cognitive processes evoked by visual external cues, i.e. CA. We evaluated the CA on the recovery of the N2 and N2pc components, while varying both the number of sweeps accounted for in the average computation (*Nswp*) and their SNR content. The *Nswp* was varied from 10–100 with a step of 10 sweeps at a time to evaluate the minimum number of sweeps required, for each SNR, to reliably estimate the N2/N2pc components.

Five standard metrics were employed to assess the performance of CA on the estimate of the N2 component, and one was employed for the assessment of the N2pc waveform: the deviation of the absolute mean around the N2 peak (*dMaP*) from a reference value, the mean absolute error in the N2 peak amplitude (AE_a) and latency (AE_l) estimates, the percentage error in the recovery of the average N2 component (E_{ave}), and the percentage error in the recovery of the average N2pc (E_{N2pc}). The mean around the peak (*MaP*) was computed as the mean value of each individual average ERP around the N2 peak position (±10 ms; peak latency estimated from noiseless GA). Then, the *dMaP* was computed for each of the 14 participants, each value of the *Nswp* (N) and each channel-target side condition as follows:

$$dMaP_i^{(N)} = \left| MaP_i{}^{(N)} - MaP_{GA} \right| , \tag{2}$$

where MaP_{GA} is the true mean around the peak computed on the GA of all the 100 individual simulated noiseless sweeps of each data group (considered as the gold standard for the error), and $MaP_i{}^{(N)}$ is the mean around the N2 peak computed for the *i*th subject using N sweeps (for $N = 10{:}10{:}100$). The absolute errors in the estimate of the N2 peak amplitude and latency were computed, for each participant and each channel-target side condition, as follows [14, 15, 53]:

$$AE_a{}^{(N)} = \left| \hat{A}_i^{(N)} - A_i \right| , \tag{3}$$

$$AE_l{}^{(N)} = \left| \hat{L}_i^{(N)} - L_i \right| , \tag{4}$$

where A_i and L_i are the true peak amplitude and latency values obtained by averaging all the 100

Table 1. Comparison between the real and simulated ranges of the N2 amplitude and latency values for the four data groups (i.e. P7R, P7L, P8R, and P8L). Bold-faced numbers indicate median values, whereas minimum and maximum values are in square brackets.

P7R		P7L		P8R		P8L	
Amplitudes (μV)							
True	*Sim*	*True*	*Sim*	*True*	*Sim*	*True*	*Sim*
−2.65	**−2.45**	**−1**	**−0.91**	**0.16**	**0.14**	**−0.33**	**−0.82**
[−4.9,2.3]	[−5.4,1.7]	[−2.9,4.0]	[−4.0,2.1]	[−3.7,2.7]	[−4.6,3.0]	[−3.4,2.7]	[−3.7,2.9]
Latencies (ms)							
True	*Sim*	*True*	*Sim*	*True*	*Sim*	*True*	*Sim*
256.9	**257.7**	**262.3**	**260.1**	**286.5**	**286.7**	**274.6**	**273.4**
[230.1,320.2]	[227.2,297.9]	[245.4,326.0]	[242.3,294.1]	[222.5,323.0]	[227.7,310.3]	[223.9,307.0]	[221.1,299.7]

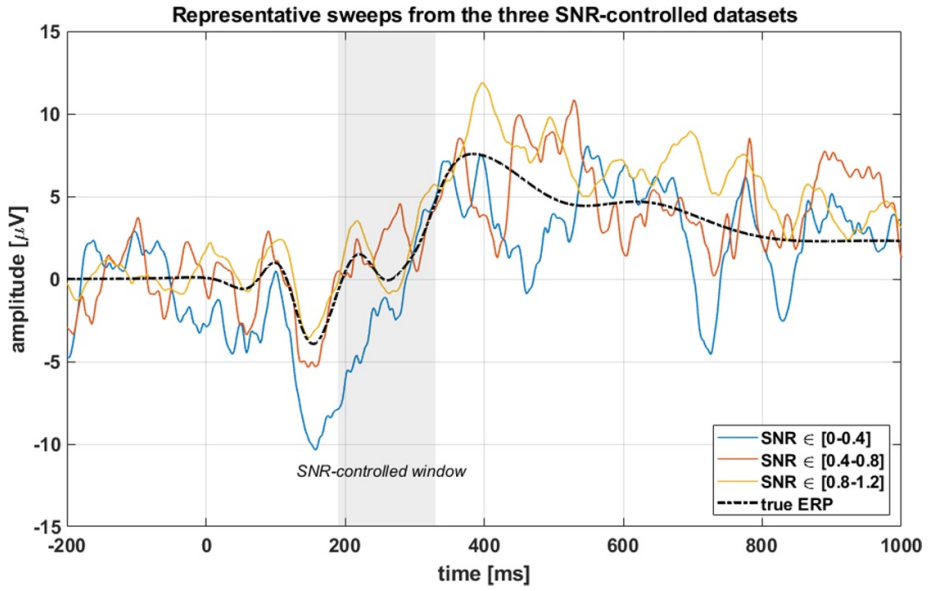

Figure 5. Three representative single noisy sweeps taken from each simulated SNR dataset; the shaded area indicates the typical N2 time window, while the black-dashed line is the noise-free synthetic signal from which the three noisy sweeps originated.

individual simulated noiseless sweeps of each data group (considered as the gold standard for the error), respectively, whereas $\hat{A}_i$ and $\hat{L}_i$ are the peak amplitude and latency estimates obtained by averaging the corresponding N noisy sweeps while varying N as previously described, respectively. The N2 peak of each noisy ERP was identified by fitting the average waveform with the corresponding GMM model to filter out the remaining noise and obtain a smoother profile. Then, the peak was identified as the local minimum in the N2 time window and checked by visual inspection.

The percentage of overall error to recover the average signal [15] was computed in the N2 time range as:

$$E_{ave}^{(N)} = \frac{||\bar{u}(t)^{(N)} - u(t)||^2}{||u(t)||^2} \times 100, \quad (5)$$

where $\bar{u}$ is the CA of N noisy sweeps, and u is the average of the 100 corresponding noise-free sweeps, i.e. the gold standard signal.

The error on the recovery of the N2pc ($E_{N2pc}^{(N)}$), obtained as the difference in the signal between CL and IL to the target activities, was additionally evaluated, as is standard in the literature [20–22]. The procedure is the same as that explained for the N2 (see equation (5)), but here the recovery of a different waveform, the N2pc, was evaluated while maintaining the same time window.

3. Results

3.1. Implementation of the simulator

3.1.1. Generation of single noise-free sweeps.

Table 1 reports, for all four groups, the median and the range of the N2 amplitude and latency values, for both the synthetic sweeps used to create the noise-free recordings and the real sweeps. A Mann–Whitney U-test did not reveal any significant statistical difference among the distributions of the N2 parameters estimated from the true vs. synthetic sweeps for each channel-target side group ($min\ p > 0.1$).

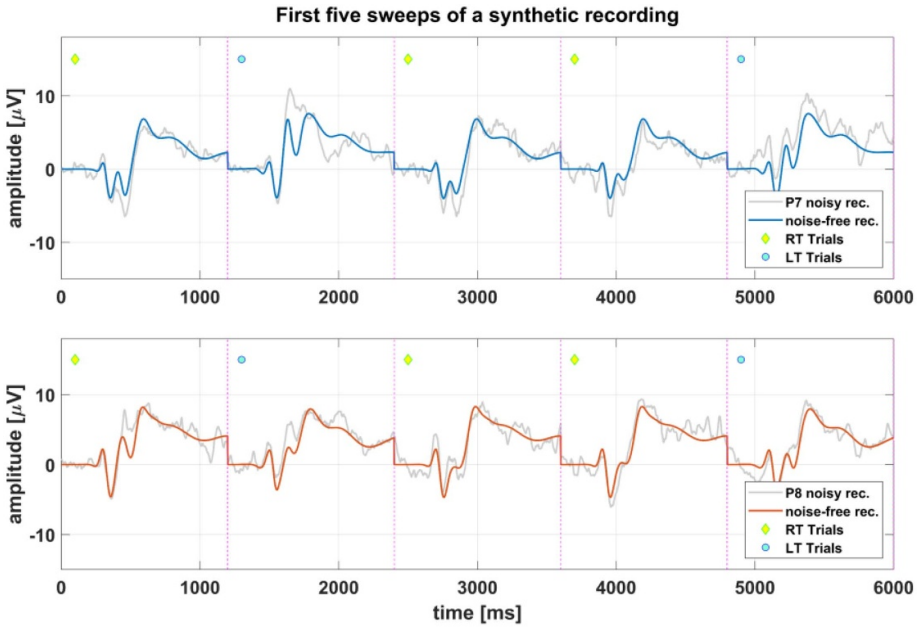

Figure 6. (Upper panel) First five noiseless sweeps (solid blue line) compared to the corresponding noisy ones (gray line) of a representative P7 synthetic recording already divided into epochs (magenta vertical lines) taken from the middle SNR dataset. Yellow diamonds indicate RT sweeps, while light blue circles the LT. Note that the main difference between the two targets is in the relative amplitude of the N2 component between P7 and P8 sweeps. (Lower panel) Representative sweeps of the corresponding P8 recording.

3.1.2. Generation of datasets with a controlled SNR.
Figure 5 shows three representative sweeps taken from each SNR dataset, superimposed to the corresponding noiseless signal (black-dashed line). The mean SNR values [meanSD] across subjects and groups for every dataset were 0.25 [0.10], 0.6 [0.12], and 0.99 [0.12], respectively. Figure 5 makes apparent, in the N2 time range, the substantial differences determined by the level of the noise between the three SNR datasets.

3.1.3. Creation of P7 and P8 noisy recordings.
Single-simulated sweeps with variability in the N2 component were arranged to create synthetic noisy recordings of 14 participants for both the P7 and P8 electrodes, with 200 sweeps each and two target conditions occurring at random. Finally, three datasets corrupted by noise but differing in their SNRs were created, as described in the previous section. Figure 6 shows the first five sweeps of P7 and P8 simulated recording (upper and lower panel, respectively) for a representative subject and SNR dataset. In each panel, the original noise-free sweeps (solid colored line) are compared with the resulting noisy sweeps in the background (gray line).

3.2. Accuracy of average N2/N2pc component estimation

3.2.1. dMaP.
The group average results of the *dMaP* are reported in figure 7, divided according to SNR (from left to right with increasing SNR range), *Nswp*, and channel-target side. The overall error values decrease significantly as the SNR range increases, and within each SNR range as the number of sweeps considered in the average increases from 10–100. In the worst SNR case, the *dMaP* is reduced from 1.3 μV (average across channel-target side groups) achieved using only 10 sweeps to 0.4 μV when using all the available sweeps. In the dataset with the SNR in the middle range, the AE_a decreases from a mean of 0.85 μV with 10 sweeps to 0.2 μV with 100, whereas in the high SNR dataset, the error decreases from 0.65–0.15 μV when considering 10 and 100 sweeps, respectively.

To individuate the minimum number of sweeps required in each SNR condition to obtain a reliable estimate of the N2, we statistically analyzed the distribution of the error across the *Nswp* and SNR ranges. Individual values of the *dMaP* were submitted to a repeated-measures analysis of variance (ANOVA) [54] with the *Nswp* (from 10–100), *Electrode* (P7 vs. P8), and *Target Side* (right vs. left) as within-subject factors, and with SNR (low, middle, and high) as a between-subjects factor. A Greenhouse–Geisser correction was applied when appropriate [55]. All statistical analyses were conducted in Rstudio [56]. The ANOVA revealed the expected significant main effects of the SNR ($F(2,39) = 29.6$, $p < .001$) and the *Nswp* ($F(9,351) = 115.9$, $p < .001$), as well as the statistically significant interaction SNR$\times$ *Nswp* ($F(18,351) = 3.3$, $p < .05$). Posthoc comparisons with a false discovery rate (FDR) control for multiple comparisons revealed statistically-significant differences among all the datasets (*min* $t(559) = 5.96$, all *ps* $< .001$).

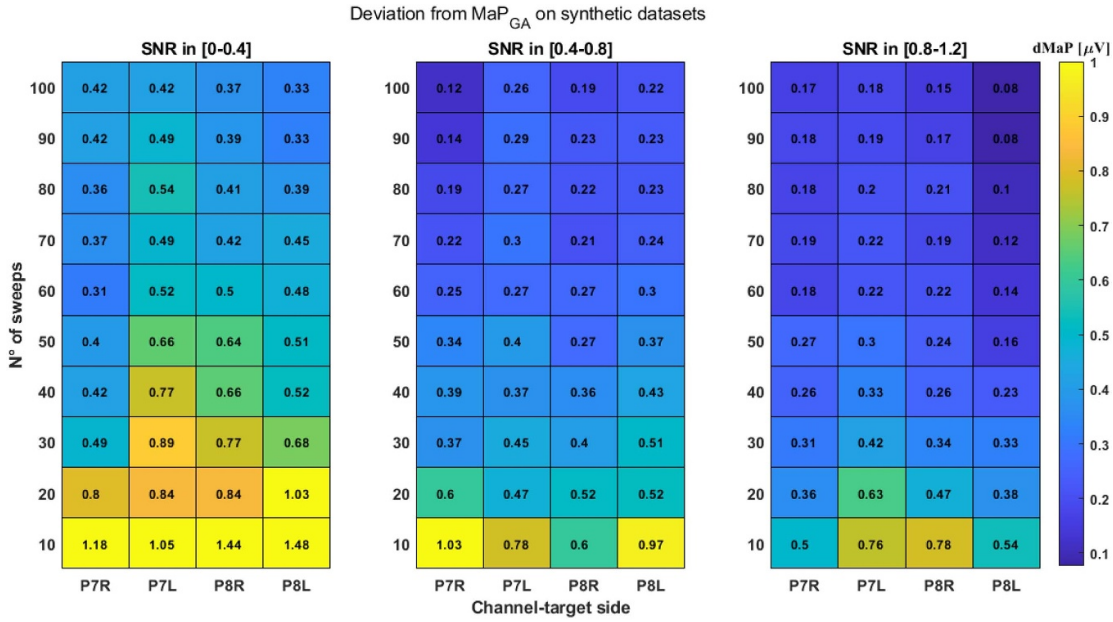

Figure 7. Colored maps of the *dMaP* for different SNR and the *Nswp* levels; values in each cell are averaged across subjects.

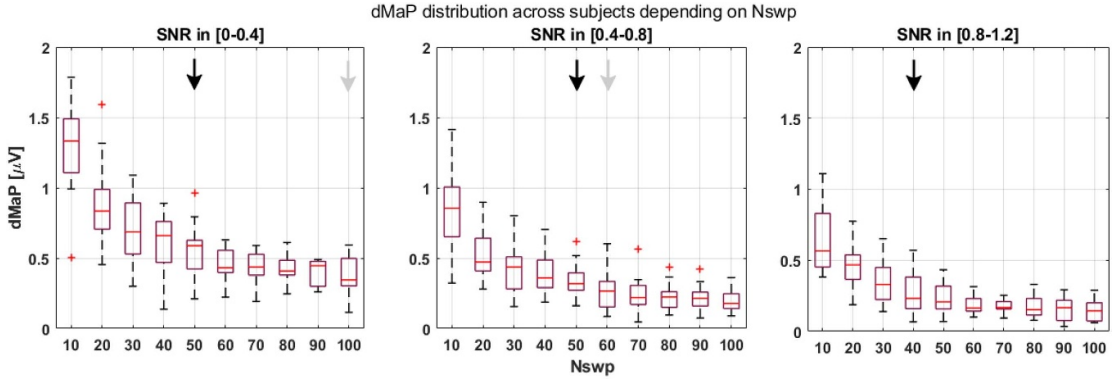

Figure 8. Distribution of the *dMaP* for different values of the *Nswp* in the three SNR datasets (panels from left to right refer to increasing SNR range, respectively). Inside each panel, the black arrow indicates the last value of the *Nswp* that was significantly different from the value of the *dMaP* at 100 trials in the same dataset, according to the t-tests. Gray arrows indicate the last value of *Nswp* in the first two datasets showing a significant difference from the error at 100 in the best SNR case. Note that for the latter, black and gray arrows were coincident.

Furthermore, to evaluate the influence of the *Nswp* on the N2 estimation accuracy, Welch's t-tests were conducted inside each SNR-controlled dataset. Data were averaged across all channel-target groups to consider only the influence of varying the *Nswp* on the overall error. The t-tests were conducted to establish when (from 10–90 sweeps) a statistically significant increase in the error value was detected, compared with the gold standard for each dataset (i.e. the average of the 100 sweeps). On the assumption that a statisticallysignificant increase in the error term implies that using fewer sweeps significantly deteriorates the estimate, we hypothesized that the level at which such a statisticallysignificant difference was detected corresponded to the minimum number of sweeps required for an N2/N2pc estimate as reliable as when having at least 100 sweeps within each SNR range. In figure 8, the boxplot representation of the error distribution across participants and channel-target groups is reported for each value of the *Nswp*, and separately for the three SNR-controlled datasets. Within each panel, the black arrow indicates the last value of the *Nswp* that still resulted in a statistically significant difference ($p < .05$) from the reference value at $Nswp = 100$ when considered inside the dataset.

The previous analysis compared the error values only within each single SNR dataset. To consider the SNR value of the dataset as well, we conducted additional t-tests following the exact same scheme described before. However, as the gold standard for all datasets, we imposed the error value at $Nswp = 100$ in the dataset with the highest SNR, which can be considered as the best-case scenario. The gray arrow in figure 8 indicates the last value of the *Nswp* that was significantly different from the value at $Nswp = 100$ in

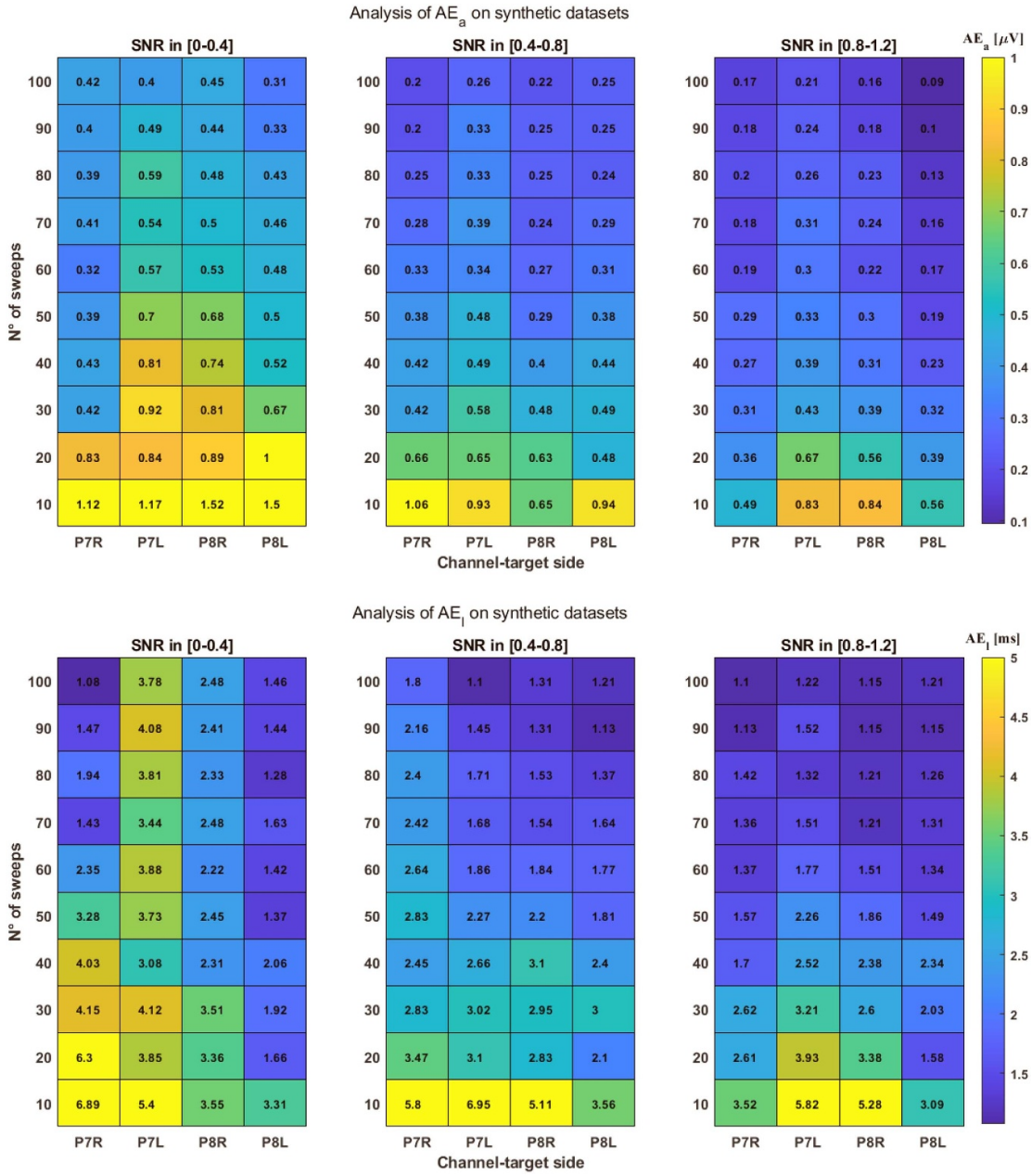

Figure 9. Colored maps of the absolute error in the estimation of the N2 peak parameters for the three SNR datasets (from left to right), each group (vertical cells) and value of the *Nswp* (horizontal cells). The top panels are related to AE_a while the bottom panels to AE_l. Values inside each cell are averaged across subjects.

the highest SNR dataset (all $ps < .05$). For example, if we consider the panel with low SNR values, the last significant difference from the corresponding average of 100 sweeps was found at 50 sweeps, but with respect to the best-case scenario, it would be at 100 sweeps, suggesting that further improvements in the estimate would be possible if the user could reduce the background noise (and therefore increase the SNR).

3.2.2. Absolute error in the N2 peak amplitude and latency estimates (AE_a and AE_l).

The statistical analysis applied to the *dMaP* was repeated for the absolute error in the estimation of the N2 peak amplitude (figure 9, top panels) and latency (figure 9, bottom panels). The results for these metrics replicated those obtained with the *dMaP*, with the overall error values showing a significant decrease as the SNR range and the *Nswp* increased. In the worst SNR case, the AE_a was reduced from 1.3 μV (average across the channel-target side groups) achieved using 10 sweeps to an average of 0.4 μV when using all the 100 available sweeps. In the dataset for the middle SNR, the AE_a decreased from a mean of 0.9 μV with 10 sweeps to a mean of 0.2 μV with 100, whereas in the best SNR dataset, the error decreased from approximately 0.7–0.2 μV. Similar conclusions could be drawn for the AE_l, where the maximum error value was, however, fairly contained at approximately 7 ms. In particular, on average across subjects, under low SNR conditions, the AE_l decreased from a mean

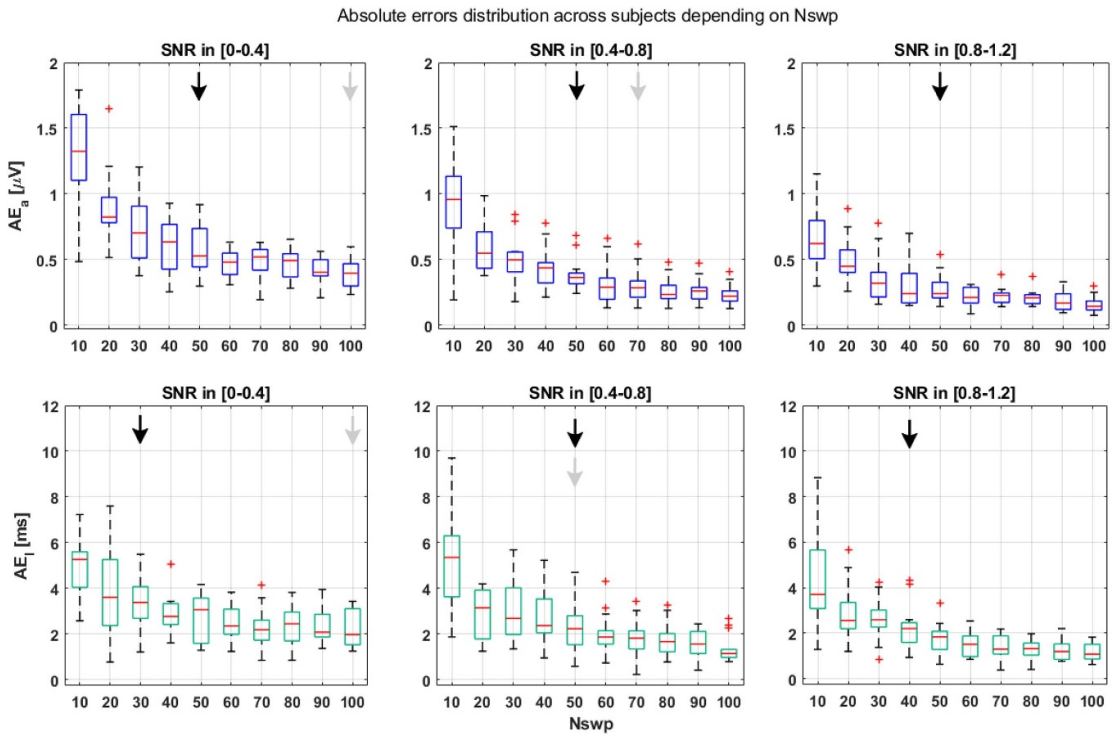

Figure 10. Distribution of AE_a (top panels) and AE_l (bottom panels) for different values of the *Nswp* and SNR range (panels from left to right refer to increasing SNR range, respectively). Within each panel, the black arrow indicates the last value of the *Nswp* revealed as statistically significant with respect to the error at *Nswp* = 100 inside the same dataset. Gray arrows indicate the last value of the *Nswp* statistically significant from the best-case scenario (i.e. the error at *Nswp* = 100 in the high SNR dataset). Note that for the high SNR, black and gray arrows were coincident.

of approximately 4.8–2.2 ms with 10 and 100 sweeps, respectively, in the middle SNR range from approximately 5.4–1.4 ms, and in the high SNR range from approximately 4.4 –1.2 ms.

Individual AE_a and AE_l values were submitted, separately, to a repeated-measures ANOVA. Similar to the case of the *dMaP*, the results revealed the expected significant main effects of the SNR ($F(2,39) = 29.97$, $p < .001$ for AE_a; $F(2,39) = 5.44$, $p < .05$ for AE_l) and *Nswp* ($F(9, 351) = 105.31$, $p < .001$ for AE_a; $F(9, 351) = 57.84$, $p < .001$ for AE_l). In addition, on the latter, a statistically significant effect was also found for the *Electrode* factor, with a higher error, on average, for P7 ($F(1,39) = 13.44$, $p < .001$). A further ANOVA performed separately for each SNR scenario revealed that this higher error for AE_l at P7 was present only in the low SNR scenario, where the ERP is highly embedded in the background noise. For AE_a, we also found statistically significant interactions for SNR × *Nswp* ($F(18,351) = 3.02$, $p < .001$) and *Electrode* × *Side* ($F(1,39) = 4.19$, $p < .05$), whereas for AE_l, the significant interactions were *Electrode* × *Side* ($F(1,39) = 5.19$, $p < .001$), *Electrode* × *Nswp* ($F(9, 351) = 3.31$, $p < .05$), and SNR × *Electrode* × *Side* × *Nswp* ($F(18,351) = 2.85$, $p < .05$).

Again, post hoc comparisons revealed statistically significant differences among all the datasets for both AE_a (*min* $t(559) = 6.75$, *all ps* $< .001$) and AE_l (*min* $t(559) = 3.06$, *all ps* $< .05$). Figure 10 reports the results of the Welch's t-test comparisons within each SNR-controlled dataset, with the aim of identifying the minimum number of sweeps yielding a statistically significant difference with the error at 100 sweeps. For AE_a, the t-test identified 60 sweeps as the minimum value to improve the error within each SNR dataset (with respect to the corresponding value of 100 sweeps). Then, further t-tests were conducted on the low and middle SNR datasets by comparing the error at each *Nswp* with the best-case scenario (namely, 100 sweeps in the high SNR dataset). The results revealed that for the low SNR case, all sweeps would be necessary to have an error comparable with that of the best case, whereas for the middle SNR case, more than 70 sweeps would be required to achieve further improvements.

The same analysis of the AE_l showed that the minimum values of the *Nswp* on the three datasets would be 40, 60, and 50, for the low, middle and high SNR, respectively, with possible improvements in the low SNR dataset only, if the SNR could be theoretically enhanced. The boxplot representation of the error distribution (considering results averaged across participants and channel-target side groups) is reported in figure 10 (upper panels for AE_a and lower panels for AE_l), for each value of the *Nswp* and the SNR.

3.2.3. Error in the recovery of the average N2 (E_{ave}). The accuracy of the average ERP in the N2 time range, estimated from the noisy recordings compared with the corresponding average noiseless ERP (i.e. the gold

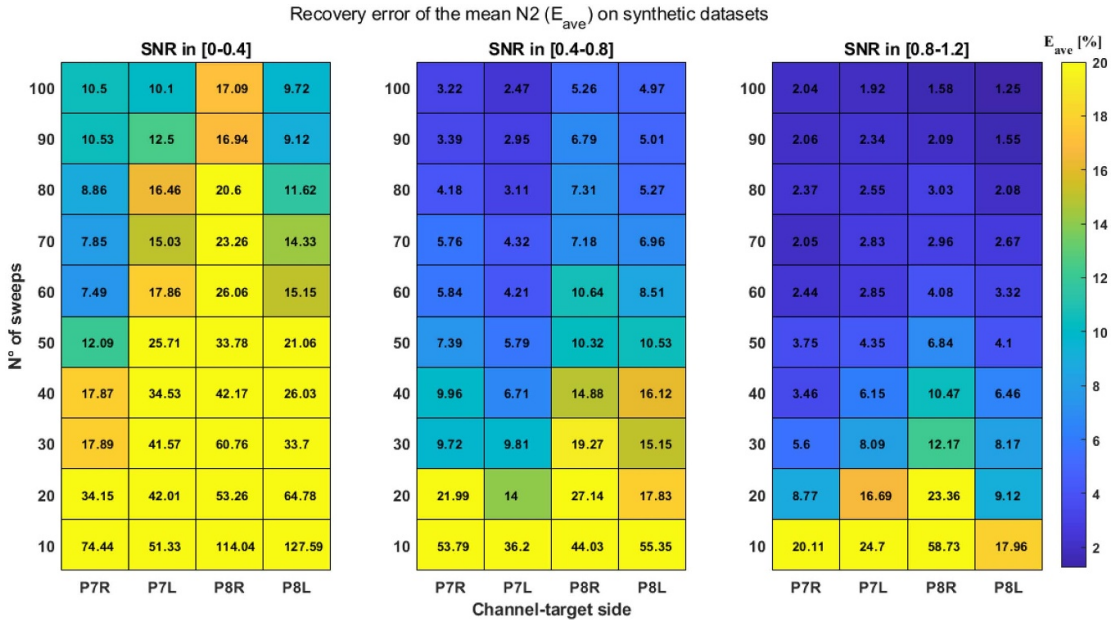

Figure 11. Colored maps of E_{ave} for different SNR and *Nswp* levels; the values in each cell are averaged across subjects.

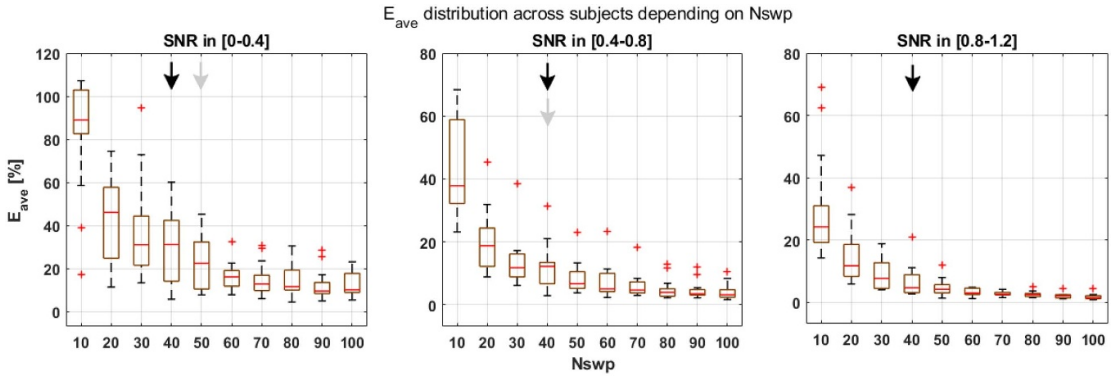

Figure 12. Distribution of E_{ave} for different values of *Nswp* and SNR (panels from left to right refer to increasing SNR range, respectively). Within each panel, the black arrow indicates the last value of the *Nswp* revealed as statistically significant with respect to the error at *Nswp* = 100 inside the same dataset. Gray arrows indicate that the last value of the *Nswp* is statistically significant from the best-case scenario (i.e. the error at *Nswp* = 100 in the high SNR dataset). Note that for the high SNR, black and gray arrows were coincidental.

standard), was additionally assessed for each value of the *Nswp* and the SNR (figure 11). The error in the recovery of the N2 component resulted in larger values in poor SNR conditions (i.e. SNR < 0.4), and when only 10 sweeps were considered in the average, with some data groups achieving E_{ave} values higher than 100%. The increase of the *Nswp* up to 100 significantly lowered the error (<15% on average across groups). In the middle SNR dataset, the error was smaller, ranging from approximately 50% with 10 sweeps to 4% with 100 sweeps (on average). In the highest SNR dataset, the error was further reduced compared with the other two datasets, ranging from approximately 30% to approximately 2% as the *Nswp* increased. The error values were submitted to a repeated-measures ANOVA. The error differed between both the SNR ($F(2,39) = 28.94$, $p < .001$) and the *Nswp* ($F(9, 351) = 82.6$, $p < .001$) factors, and post hoc comparisons of SNR values revealed significant differences among all datasets (*min* $t(559) = 6.61$, all *ps* < .001). Furthermore, the interactions of SNR × *Nswp* and *Electrode* × *Nswp* were significant ($F(18,351) = 7.94$, $p < .001$ and $F(9, 351) = 3.31$, $p < .05$, respectively). Welch's t-tests were applied to each SNR dataset to investigate whether significant differences across errors were present in the *Nswp* levels. Here, the first t-test identified the value of 50 as the most conservative *Nswp* value to obtain a significant improvement with respect to the highest *Nswp* value in each dataset. The second t-test, which used the gold standard in the best SNR scenario, revealed the values of *Nswp* = 60 and *Nswp* = 50 in the low and the middle SNR datasets, respectively (figure 12).

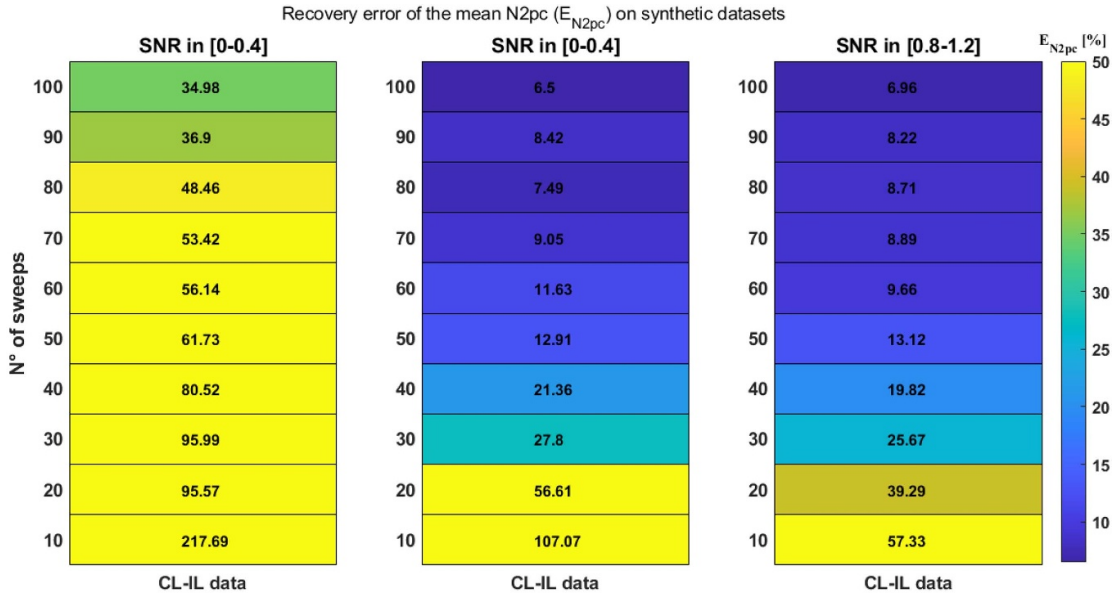

Figure 13. Colored maps of E_{N2pc} for different SNR ranges and *Nswp* levels; the values in each cell are averaged across subjects.

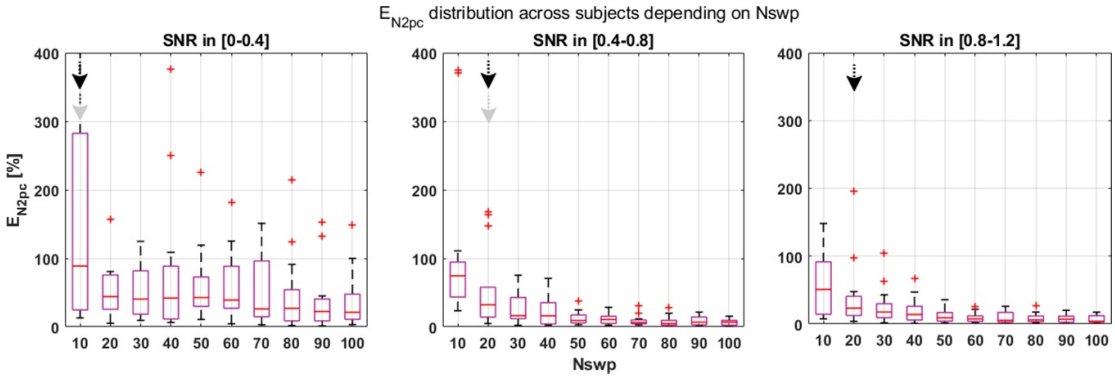

Figure 14. Distribution of E_{N2pc} varying the *Nswp* for the three SNR datasets (panels from left to right refer to increasing SNR range, respectively). Black arrows indicate the last value of the *Nswp* that was significantly different from the value of E_{N2pc} at 100 trials in the same dataset, according to the t-test. Gray arrows indicate the last value of the *Nswp* in the first two datasets that was significantly different from the error at 100 in the high SNR dataset. Note that for the latter, black and gray arrows would be coincident.

3.2.4. Error in the recovery of the average N2pc (E_{N2pc}). For the sake of completeness, the same estimate of error obtained in the recovery of the average N2 was computed for the difference in waveform, i.e. the N2pc, obtained from the subtraction between the average CL and IL N2 signals. The global results are shown in figure 13, where, within each SNR-related panel, the vertical cell refers to the CL/IL difference computed between the P7/P8 channels according to target position. For the N2pc, the recovery error reduces as the SNR of the data in the N2 time range improves and the *Nswp* increases, with higher error values observed in the lower SNR dataset and with 10 sweeps, as expected. In the lower SNR dataset, the error on average across subjects decreases from approximately 200% to approximately 35% with 10 and 100 sweeps, respectively, whereas the error in the middle SNR dataset decreases from approximately 100%–6.5%, and in the highest SNR dataset from approximately 60%–7%.

The ANOVA was applied to the data with the *Nswp* (from 10–100) as the within-subject factor and the SNR (low, middle, and high) as the between-subject factor. The results revealed significant differences in the data for the SNR ($F(2,39) = 5.77$, $p < .05$) and *Nswp* ($F(9, 351) = 10.28$, $p < .001$) factors. Post hoc comparisons with an FDR adjustment revealed significant differences between the low and middle SNR datasets ($t(139) = 4.2$, $p < .001$) and between the low and high SNR datasets ($t(139) = 5.2$, $p < .001$), whereas no significant difference emerged between the middle and high SNR cases. Additional t-tests were conducted within each dataset to find at which *Nswp* there was a significant increase in the recovery error compared to the gold standard. The black arrows in the boxplot representation of figure

14 point at the results of the Welch's t-tests inside the same dataset, whereas the gray arrows indicate the results of the second round of t-tests that identified further possible error improvements with respect to the best SNR case.

3.2.5. Generalization of the outcomes of the entire sample of subjects.

To verify whether the distribution of each error metric at the different *Nswp* was homogeneous among subjects and confirm the generalization of the outcomes to the entire sample, Levene's test on the equality of variances was applied to the outcomes [57]. The procedure was the same as the Welch t-tests whereby a comparison among error distributions corresponding to $Nswp < 100$ and the reference at $Nswp = 100$ was performed for each error metric. If the variances were found to be unequal, the validity of the recommended *Nswp* could be affected. Levene's tests showed that inside each individual SNR dataset—for all metrics and SNR scenarios except *dMaP* in the low SNR—only the error variance at $Nswp = 10$ and/or 20 was significantly different ($p < .05$) from the reference (1st Levene's test). When comparing errors against the best SNR scenario (2nd Levene's test), for the low SNR dataset, AE_a and E_{ave} showed error distributions with significantly different variance ($p < .05$) from the reference at $Nswp = 10$, 20, and 30, AE_l and *dMaP* at $Nswp = 10$ and 20, and E_{N2pc} at $Nswp = 10$. For the middle SNR, instead, $Nswp = 10$ for all metrics, and $Nswp = 20$ for E_{ave} and E_{N2pc}, were different from the reference variance. These results suggest that the error distributions corresponding to the minimum *Nswp* recommended for each metric and SNR are homogeneous across the 14 subjects and, therefore, that our results could be generalized to the entire sample.

3.2.6. Summary of the main results.

Table 2 summarizes the minimum values of the *Nswp* as a function of the tested metric and SNR scenario. The top three rows of table 2 report the results of the 1st Welch's t-test when the error distributions at each $Nswp < 100$ and $Nswp = 100$, inside each SNR scenario and error metric, were compared. The bottom three rows of the table report the results of the 2nd Welch's t-test when error distributions at $Nswp < 100$ were compared to the best-case scenario, namely, $Nswp = 100$ in the high SNR dataset. Not surprisingly, the results for the best SNR scenario are identical in the two t-tests because this scenario was taken as a reference in the 2nd t-test. When collectively taken, the results show that the minimum number of sweeps differs among error metrics, being particularly low for the N2pc (i.e. about 20–30 sweeps, depending on the SNR of the data). This could be due to the fact that the N2pc, that is a difference in ERP component resulting from the subtraction of two N2 waves, is totally embedded in the spontaneous EEG activity and its reliable detection is definitely more challenging than the N2 itself, independently of the *Nswp*. In this vein, the results suggest that, for the N2pc, it is not worth acquiring more than 20–30 sweeps in an experiment with comparable conditions. Exceeding this value, in practice, does not entail any substantial improvement in the final estimate of the average ERP component. Depending on the aim of their investigations, users should therefore choose an *Nswp* based on their will to control a specific error or make a trade-off among them.

Table 2. Summary of the main results for each error metric and SNR scenario. The values reported in the table represent the lower bound of the *Nswp* to achieve a statistically significant lower error with respect to the reference value used for each Welch's t-test comparison.

	SNR range	*dMaP*	AE_a	AE_l	E_{ave}	E_{N2pc}
1st t-test	SNR $\in$ [0.1–0.4]	>50	>50	>30	>40	>10
	SNR $\in$ [0.4–0.8]	>50	>50	>50	>40	>20
	SNR $\in$ [0.8–1.2]	>40	>50	>40	>40	>20
2nd t-test	SNR $\in$ [0.1–0.4]	>100	>100	>100	>50	>10
	SNR $\in$ [0.4–0.8]	>60	>70	>50	>40	>20
	SNR $\in$ [0.8–1.2]	>40	>50	>40	>40	>20

4. Discussion

The main result of the present work is that, although the SNR strongly influences the decision on the minimum number of necessary sweeps for a given SNR (as expected), no significant improvements in the estimate of the N2 and N2pc could be achieved, in each SNR scenario, beyond a certain value of *Nswp* (denoted by the black arrows on the boxplots in figures 8, 10, 12, and 14). This result suggests that background noise is not further suppressed by increasing the *Nswp* beyond the minimum values reported herein. This should allow researchers to reduce the length of their experiments and the influence of additional (and subject-dependent) noise sources that could typically arise when long experiments are carried out, e.g. fatigue. However, if researchers wish to test this experimental circumstance, a manipulation of the SNR could be employed to generate synthetic recordings accounting for fatigue effects. Intuitively, when fatigue emerges, the SNR of the data should worsen throughout the experiment because of a decrease in the ERP power. Modeling the decline of the SNR over time could be a promising avenue for a further refinement of the presently proposed simulator.

Our results should guide the data acquisition process when CA is going to be applied to estimate the N2/N2pc components. The minimum *Nswp* may occur very early, as in the case of low SNR, highlighting that when the SNR of the signal is very poor, increasing the *Nswp* has a negligible effect on the quality of the estimate. This is mainly because of the high variability among the estimates. In

contrast, in high and middle SNR conditions, significant enhancements in the N2 and N2pc estimate could sometimes still be visible when increasing the *Nswp* above the best value obtained for the worst SNR case (as, for example, occurs for AE_l). Although this result might be counter-intuitive, it is not if we hypothesize that, at a high SNR, the variability among estimates is highly reduced compared with lower SNR conditions, thus facilitating the statistical outcome.

Our investigation also highlighted that the minimum value of the *Nswp* may differ among the *dMaP*, AE_a, AE_l, E_{ave}, and E_{N2pc}. Therefore, in practice, users should decide whether to estimate the number of sweeps with a trade-off between the errors or select which one to control the most, depending on the specific application and desired metric for their analysis. Users should be aware that, if averaging is used, as it happens in most of the real applications, each error metric contains two contributions: the amount of suppression of the background noise and the intra-individual fluctuations of the N2 across sweeps. For E_{ave} the latter tends to 0 as the *Nswp* increases. In our simulation, the mean (±standard deviation) oscillation of the noise-free average signal in the N2 time window and considering $Nswp = 50$ (the minimum value of *Nswp* obtained for E_{ave} in each SNR dataset) was lower than 1.04% (±0.54%) of the gold-standard value, corresponding to $Nswp = 100$, on average across subjects and data groups. The contribution of N2 oscillations across sweeps to the error metric E_{ave} was of about 4.4%, 12.3%, and 20.8% in the low, middle and high SNR scenarios, respectively, on average across subjects and data groups. These results highlight that the larger contribution to the error metric is due to the background noise.

The present study also investigated differences in the initial noise content of the data. Therefore, it is important to remark that, in practice, users should estimate the likely SNR of their data before selecting the number of sweeps to be implemented in their experiment. In our simulation context, it was possible to easily and reliably estimate and control the SNR of the synthetic sweeps in the N2 time range, from which we generated three datasets with different and plausible SNR ranges based on results on the available real data and on previous studies [14, 15, 50, 53], as described in section 2.1.6. Although it might seem difficult to estimate the SNR of data without a noise-free signal available and before acquiring the data, some approaches exist in the literature [52, 58, 59]. It should, however, be noted that here the SNR was computed only in the N2 time range, which was the range of interest, and therefore the SNR of the entire sweep could be outside of the desired SNR range.

Another point that also deserves discussion is that, in the present manuscript, in generating the noise of the synthetic sweeps, we used real RS data (ad hoc manipulated to generate different noise contexts). In the absence of this kind of RS data, several approaches including autoregressive (AR) models, adaptive Markov process amplitude (AMPA) algorithms, and artificial neural networks (ANNs) could be used to simulate realistic background EEG noise. In the future, one of these methods could be integrated into our simulator to generate random and fully synthetic EEG signals. For example, simple AR models could be employed to generate background EEG noise, provided that a short interval of data (e.g. the baseline) is preliminarily used to identify the model parameters [14, 60]. AMPA, expressed by sinusoidal waves, could be used to model and simulate non-stationary EEG (e.g. in the presence of artefacts) [61]. Finally, ANNs are valid tools for generating long-lasting EEG signals (e.g. those of channels P7 and P8 used in our study to compute the N2/N2pc components) through an iterative simulation process, without the use of pre-recorded EEG sequences [62, 63].

5. Conclusions

The choice of the number of sweeps to be presented to participants in an ERP experiment is a complex and multifaceted issue, which should consider, among other factors, the noise content of the signal and the magnitude of the component of interest. Realistic simulation scenarios could help users set the number of sweeps for an experiment in an objective way, while manipulating the aforementioned factors according to their specific experimental setting. Here, we implemented a flexible and easily tuneable simulation framework and demonstrated one among several possible applications, i.e. the evaluation of the influence of the SNR of the signal and of the number of averaged sweeps on the accuracy of the estimate of the N2/N2pc components with the standard averaging method. The advantage of the simulator described in this work is that it is easily reproducible, generalizable, and flexible. Its tuneable parameters allow users to mimic and investigate several experimental conditions, e.g. different SNR ranges, fatigue effects, unexpected cases, as well as ERP components other than N2/N2pc using different experimental protocols. Furthermore, the presented simulator for N2/N2pc data could be a useful tool for more challenging tasks, such as the development of techniques for single-trial estimation to investigate the variability of these components throughout the experiment [15, 64–66].

In conclusion, the simulator and the implementation documented in the present paper are potentially useful to design novel experiments aiming to study attention processes expressed by N2/N2pc modulations. The most suitable number of sweeps should reflect the specific type of error to minimize. The tables reported in the paper provide guidelines for the setup of these novel paradigms. Furthermore, the simulator developed in this paper will be extremely useful to other neuroscientists willing to validate their

novel processing/estimation techniques on synthetic data emulating the variability and noise content of real EEG data.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

ORCID iDs

Francesca Marturano https://orcid.org/0000-0002-5430-1234
Sabrina Brigadoi https://orcid.org/0000-0003-3032-7381
Mattia Doro https://orcid.org/0000-0003-2574-4308
Roberto Dell'Acqua https://orcid.org/0000-0002-3393-1907
Giovanni Sparacino https://orcid.org/0000-0002-3248-1393

References

[1] Hämäläinen J A, Salminen H K and Leppänen P H T 2013 Basic auditory processing deficits in dyslexia *J. Learn. Disabil.* **46** 413–27
[2] De Tommaso M, Navarro J, Ricci K, Lorenzo M, Lanzillotti C, Colonna F, Resta M, Lancioni G and Livrea P 2013 Pain in prolonged disorders of consciousness: laser evoked potentials findings in patients with vegetative and minimally conscious states *Brain Inj.* **27** 962–72
[3] Ethridge L E, White S P, Mosconi M W, Wang J, Byerly M J and Sweeney J A 2016 Reduced habituation of auditory evoked potentials indicate cortical hyper-excitability in Fragile X Syndrome *Transl. Psychiatry* **6** e787
[4] Kang E, Keifer C M, Levy E J, Foss-Feig J H, McPartland J C and Lerner M D 2018 Atypicality of the N170 event-related potential in autism spectrum disorder: a meta-analysis *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3** 657–66
[5] Tonin L, Leeb R and Del Millán R J 2012 Time-dependent approach for single trial classification of covert visuospatial attention *J. Neural Eng.* **9** 045011
[6] Treder M S, Bahramisharif A, Schmidt N M, Van Gerven M A and Blankertz B 2011 Brain-computer interfacing using modulations of alpha activity induced by covert shifts of attention *J. Neuroeng. Rehabil.* **8** 24
[7] Tian Y, Zhang H, Li P and Li Y 2019 Multiple correlated component analysis for identifying the bilateral location of target in visual search tasks *IEEE Access* **7** 98486–94
[8] Mazza V and Caramazza A 2011 Temporal brain dynamics of multiple object processing: the flexibility of individuation, ed S He *PLoS One* **6** e17453
[9] Benavides-Varela S, Basso M S, Brigadoi S, Meconi F, Doro M, Simion F, Sessa P, Cutini S and Dell'Acqua R 2018 N2pc reflects two modes for coding the number of visual targets *Psychophysiology* **55** e13219
[10] Lotte F 2011 Generating artificial EEG signals to reduce BCI calibration time *5th Int. Brain-Comp. Interface Workshop* pp 176–9
[11] Delorme A, Mullen T, Kothe C, Akalin Acar Z, Bigdely-Shamlo N, Vankov A and Makeig S 2011 EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing *Comput. Intell. Neurosci.* **2011** 130714
[12] Lindgren J T, Merlini A, Lecuyer A and Andriulli F P 2018 simBCI—a framework for studying BCI methods by simulated EEG *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 2096–105
[13] Kiesel A, Miller J, Jolicœur P and Brisson B 2008 Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods *Psychophysiology* **45** 250–74
[14] D'Avanzo C, Schiff S, Amodio P and Sparacino G 2011 A Bayesian method to estimate single-trial event-related potentials with application to the study of the P300 variability *J. Neurosci. Methods* **198** 114–24
[15] D'Avanzo C, Goljahani A, Pillonetto G, De Nicolao G and Sparacino G 2013 A multi-task learning approach for the extraction of single-trial evoked potentials *Comput. Methods Programs Biomed.* **110** 125–36
[16] Oostenveld R, Fries P, Maris E and Schoffelen J M 2011 FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data *Comput. Intell. Neurosci.* **2011** 156869
[17] Tadel F, Baillet S, Mosher J C, Pantazis D and Leahy R M 2011 Brainstorm: a user-friendly application for MEG/EEG analysis *Comput. Intell. Neurosci.* **2011** 879716
[18] Haufe S and Ewald A 2019 A simulation framework for benchmarking eeg-based brain connectivity estimation methodologies *Brain Topogr.* **32** 625–42
[19] Tan M and Wyble B 2015 Understanding how visual attention locks on to a location: toward a computational model of the N2pc component *Psychophysiology* **52** 199–213
[20] Eimer M 1996 The N2pc component as an indicator of attentional selectivity *Electroencephalogr. Clin. Neurophysiol.* **99** 225–34
[21] Corriveau I, Fortier-Gauthier U, Pomerleau V J, McDonald J, Dell'Acqua R and Jolicoeur P 2012 Electrophysiological evidence of multitasking impairment of attentional deployment reflects target-specific processing, not distractor inhibition *Int. J. Psychophysiol.* **86** 152–9
[22] Pomerleau V J, Fortier-Gauthier U, Corriveau I, McDonald J J, Dell'Acqua R and Jolicœur P 2014 The attentional blink freezes spatial attention allocation to targets, not distractors: evidence from human electrophysiology *Brain Res.* **1559** 33–45
[23] Hickey C, Di Lollo V and McDonald J J 2009 Electrophysiological indices of target and distractor processing in visual search *J. Cogn. Neurosci.* **21** 760–75
[24] Luck S J and Hillyard S A 1994 Spatial filtering during visual search: evidence from human electrophysiology *J. Exp. Psychol.: Hum. Percept. Perform.* **20** 1000–14
[25] Brisson B and Jolicœur P 2007 The N2pc component and stimulus duration *Neuroreport* **18** 1163–6
[26] Woodman G F and Luck S J 2003 Serial deployment of attention during visual search *J. Exp. Psychol. Hum. Percept. Perform.* **29** 121–38
[27] Kiss M, Van Velzen J and Eimer M 2008 The N2pc component and its links to attention shifts and spatially selective visual processing *Psychophysiology* **45** 240–9
[28] Tonin L, Leeb R, Sobolewski A and Del R Millán J 2013 An online EEG BCI based on covert visuospatial attention in absence of exogenous stimulation *J. Neural Eng.* **10** 056007
[29] Woodman G F, Arita J T and Luck S J 2009 A cuing study of the N2pc component: an index of attentional deployment to objects rather than spatial locations *Brain Res.* **1297** 101–11
[30] Möckel T, Beste C and Wascher E 2015 The effects of time on task in response selection—an ERP study of mental fatigue *Sci. Rep.* **3** 10113
[31] Larson M J, Baldwin S A, Good D A and Fair J E 2010 Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials *Psychophysiology* **47** 1167–71
[32] Fischer A G, Klein T A and Ullsperger M 2017 Comparing the error-related negativity across groups: the impact of error- and trial-number differences *Psychophysiology* **54** 998–1009
[33] Duncan C C, Barry R J, Connolly J F, Fischer C, Michie P T, Näätänen R, Polich J, Reinvang I and Van Petten C 2009

Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400 *Clin. Neurophysiol.* **120** 1883–908
[34] Marco-Pallares J, Cucurell D, Münte T F, Strien N and Rodriguez-Fornells A 2011 On the number of trials needed for a stable feedback-related negativity *Psychophysiology* **48** 852–60
[35] Boudewyn M A, Luck S J, Farrens J L and Kappenman E S 2018 How many trials does it take to get a significant ERP effect? It depends *Psychophysiology* **55** e13049
[36] Klem G H, Lüders H O, Jasper H H and Elger C 1999 The ten-twenty electrode system of the International Federation. The International Federation of clinical neurophysiology *Electroencephalogr. Clin. Neurophysiol. Suppl.* **52** 3–6
[37] Lakshmi M R, Prasad T V and Chandra Prakash V 2014 Survey on EEG signal processing methods *Int. J. Adv. Res. Comp. Sci. Softw. Eng.* **4** 84
[38] Delorme A and Makeig S 2004 EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis *J. Neurosci. Methods* **134** 9–21
[39] Jolicœur P, Brisson B and Robitaille N 2008 Dissociation of the N2pc and sustained posterior contralateral negativity in a choice response task *Brain Res.* **1215** 160–72
[40] Göddertz A, Klatt L-I, Mertes C and Schneider D 2018 Retroactive attentional shifts predict performance in a working memory task: evidence by lateralized EEG patterns *Front. Hum. Neurosci.* **12** 428
[41] Awni H, Norton J J S, Umunna S, Federmeier K D and Bretl T 2013 Towards a brain computer interface based on the N2pc event-related potential *Int. IEEE/EMBS Conf. on Neural Eng.* pp 1021–4
[42] Praamstra P and Kourtis D 2010 An early parietal ERP component of the frontoparietal system: EDAN ≠ N2pc *Brain Res.* **1317** 203–10
[43] Ai G, Sato N, Singh B and Wagatsuma H 2016 Direction and viewing area-sensitive influence of EOG artifacts revealed in the EEG topographic pattern analysis *Cogn. Neurodyn.* **10** 301–14
[44] Jia Y and Tyler C W 2019 Measurement of saccadic eye movements by electrooculography for simultaneous EEG recording *Behav. Res. Methods* **51** 2139–51
[45] Ting C-M, Salleh S-H, Zainuddin Z M and Bahar A 2015 Modeling and estimation of single-trial event-related potentials using partially observed diffusion processes *Digital Signal Process.* **36** 128–43
[46] Mortaheb S, Rostami F, Shahin S and Amirfattahi R 2016 Wavelet based single trial event related potential extraction in very low SNR conditions *2016 6th Int. Conf. on Computer and Knowledge Engineering (ICCKE)* pp 82–7
[47] Luck S 2005 *An Introduction to the Event-Related Potential Technique* (Cambridge, MA: MIT Press)
[48] Luck S J and Kappenman E S 2012 *Oxford Handbook of Event-Related Potential Components* (New York: Oxford University Press) (https://doi.org/10.1093/oxfordhb/9780195374148.001.0001)
[49] Akaike H 1974 A new look at the statistical model identification *IEEE Trans. Autom. Control* **19** 716–23
[50] Zouridakis G, Iyer D, Diaz J and Patidar U 2007 Estimation of individual evoked potential components using iterative independent component analysis *Phys. Med. Biol.* **52** 5353–68
[51] Gonen F F and Tcheslavski G V 2012 Techniques to assess stationarity and gaussianity of EEG: an overview *Int. J. Bioautom.* **16** 135–42
[52] Elberling C and Don M 1984 Quality estimation of averaged auditory brainstem responses *Scand. Audiol.* **13** 187–97
[53] Fukami T, Watanabe J and Ishikawa F 2016 Robust estimation of event-related potentials via particle filter *Comput. Methods Programs Biomed.* **125** 26–36
[54] Girden E R 1992 *ANOVA: Repeated Measures* (Newbury Park, CA: Sage Publications) (https://doi.org/10.4135/9781412983419)
[55] Greenhouse S W and Geisser S 1959 On methods in the analysis of profile data *Psychometrika* **24** 95–112
[56] RStudio Team 2015 *RStudio: Integrated Development Environment for R* Software v0.98.1074 (Boston, MA: RStudio) www.rstudio.com
[57] Gastwirth J L, Gel Y R and Miao W 2009 The impact of Levene's test of equality of variances on statistical theory and practice *Stat. Sci.* **24** 343–60
[58] Silva I 2009 Estimation of postaverage SNR from evoked responses under nonstationary noise *IEEE Trans. Biomed. Eng.* **56** 2123–30
[59] Hu L, Mouraux A, Hu Y and Iannetti G D 2010 A novel approach for enhancing the signal-to-noise ratio and detecting automatically event-related potentials (ERPs) in single trials *Neuroimage* **50** 99–111
[60] Marturano F, Brigadoi S, Doro M, Roberto D and Sparacino G 2019 Development of a computer simulator of the visual N2 event-related potential component for the study of cognitive processes *XV Mediterranean Conf. on Medical and Biological Engineering and Computing* **79** 29–36
[61] Al-Nashash H, Al-Assaf Y, Paul J and Thakor N 2004 EEG signal modeling using adaptive Markov process amplitude *IEEE Trans. Biomed. Eng.* **51** 744–51
[62] Tomasevic N M, Neskovic A M and Neskovic N J 2012 Artificial neural network based approach to EEG signal simulation *Int. J. Neural Syst.* **22** 1250008
[63] Tomasevic N M, Neskovic A M and Neskovic N J 2017 Correlated EEG signals simulation based on artificial neural networks *Int. J. Neural Syst.* **27** 1750008
[64] Manresa J A B, Arguissain F G, Redondo D E M, Mørch C D and Andersen O K 2015 On the agreement between manual and automated methods for single-trial detection and estimation of features from event-related potentials *PLoS One* **10** e0134127
[65] Lee W L, Tan T, Falkmer T and Leung Y H 2016 Single-trial event-related potential extraction through one-unit ICA-with-reference *J. Neural Eng.* **13** 066010
[66] Blankertz B, Lemm S, Treder M, Haufe S and Müller K R 2011 Single-trial analysis and classification of ERP components—a tutorial *Neuroimage* **56** 814–25