International Journal of Psychophysiology 144 (2019) 14–24

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Full Length Article

Single-subject analysis of N400 event-related potential component with five different methods

Roosa E. Kallionpää[a,b,*], Henri Pesonen[c,d], Annalotta Scheinin[b,e], Nils Sandman[a], Ruut Laitio[e], Harry Scheinin[b,e,f], Antti Revonsuo[a,g], Katja Valli[a,b,g]

[a] *Department of Psychology and Speech-Language Pathology, and Turku Brain and Mind Center, University of Turku, Turku, Finland*
[b] *Department of Perioperative Services, Intensive Care and Pain Medicine, Turku University Hospital, Turku, Finland*
[c] *Department of Mathematics and Statistics, University of Turku, Turku, Finland*
[d] *Department of Computer Science, Aalto University, Espoo, Finland*
[e] *Turku PET Centre, University of Turku and Turku University Hospital, Turku, Finland*
[f] *Integrative Physiology and Pharmacology, Institute of Biomedicine, University of Turku, Turku, Finland*
[g] *Department of Cognitive Neuroscience and Philosophy, University of Skövde, Skövde, Sweden*

ARTICLE INFO

Keywords:
Data analysis
Electroencephalography
Event-related potentials
N400 evoked potential
Statistical data interpretation

ABSTRACT

There are several different approaches to analyze event-related potentials (ERPs) at single-subject level, and the aim of the current study is to provide information for choosing a method based on its ability to detect ERP effects and factors influencing the results. We used data from 79 healthy participants with EEG referenced to mastoid average and investigated the detection rate of auditory N400 effect in single-subject analysis using five methods: visual inspection of participant-wise averaged ERPs, analysis of variance (ANOVA) for amplitude averages in a time window, cluster-based non-parametric testing, a novel Bayesian approach and Studentized continuous wavelet transform (t-CWT). Visual inspection by three independent raters yielded N400 effect detection in 85% of the participants in at least one paradigm (active responding or passive listening), whereas ANOVA identified the effect in 68%, the cluster-method in 59%, the Bayesian method in 89%, and different versions of t-CWT in 22–59% of the participants. Thus, the Bayesian method was the most liberal and also showed the greatest concordance between the experimental paradigms (active/passive). ANOVA detected significant effect only in cases with converging evidence from other methods. The t-CWT and cluster-based method were the most conservative methods. As we show in the current study, different analysis methods provide results that do not completely overlap. The method of choice for determining the presence of an ERP component at single-subject level thus remains unresolved. Relying on a single statistical method may not be sufficient for drawing conclusions on single-subject ERPs.

1. Introduction

Single-subject analyses of event-related potentials (ERPs) and related statistical testing have been widely discussed due to contradictory results reported for different approaches (Gabriel et al., 2016; Groppe et al., 2011b), incomplete reproducibility (Naccache et al., 2015; Tzovara et al., 2015) and biased chasing for significance (Luck and Gaspelin, 2017). Further, in addition to the fact that there is no explicit observational definition of an ERP component, the methods used in single-subject analyses are an inconsistent mixture, varying in a priori information, statistical corrections, and clinical conventions. This makes the direct comparison of the results produced by the different analysis methods unfeasible. Yet, analysis at single-subject level is needed when ERPs are used to elucidate individual characteristics such as person's state of consciousness or the effect of a neurological disorder.

In clinical setting, it is usually sufficient to find out whether the ERPs produced by two types of stimuli differ in terms of amplitude, or whether a certain ERP component is present. This has traditionally been assessed by visual inspection, despite the risk of subjectivity. To increase reliability, many researchers have used multiple methods for single-subject analyses, but the results have overlapped only partially (Rohaut et al., 2015; Sculthorpe-Petley et al., 2015; Steppacher et al., 2013). An ideal statistical method for single-subject analyses would be sensitive but reliable in terms of false positives, objective, easy to use, and utilizable with many ERP components. The method should also

[*] Corresponding author at: Department of Psychology and Speech-Language Pathology, University of Turku, Assistentinkatu 7, 20014 Turku, Finland.
E-mail address: roosa.kallionpaa@utu.fi (R.E. Kallionpää).

https://doi.org/10.1016/j.ijpsycho.2019.06.012
Received 22 January 2019; Received in revised form 13 June 2019; Accepted 18 June 2019
Available online 19 June 2019

take advantage of all the data, function without broad a priori assumptions, and be able to handle the multiple comparisons problem.

Single-subject analyses are also complicated by differences in ERP amplitude, latency and scalp distribution between individuals (Lang et al., 1995; Luck et al., 2011), between healthy subjects and patients with brain injuries (Duncan et al., 2009; Kotchoubey, 2015), and even between different recordings from the same person (Lang et al., 1995). The limited number of trials and the resulting signal-to-noise ratio (SNR) also hinder the use of single-subject level analyses. Especially when complex stimuli are necessary, such as with cognitive ERPs, the duration of an experiment and the laborious preparation of stimuli often limit the number of trials. This makes the choice of analysis method especially crucial.

Our general aim with this study is to provide information to both researchers and clinicians to help them make informed decisions on the selection of single-subject analysis methods, and to compare and interpret results of ERP studies. Specifically, we used EEG data from a large experiment and studied N400 ERP component to identify, elucidate, and substantiate the effects of five different single-subject analysis methods on the detection of a complex cognitive ERP. In addition to visual inspection, we utilized analysis of variance (ANOVA) of amplitude averages in a time window, cluster-based non-parametric method, a Bayesian approach, and Studentized continuous wavelet transform (t-CWT). The basic principles and characteristics of these methods are introduced in the following paragraphs.

1.1. Visual inspection

Visual inspection of averaged ERPs allows taking into account individual differences in topography, latency and morphology of ERP components, as well as interactions with other components. This flexibility of the method is especially important when, e.g., patients with delayed ERP latencies are studied. Visual inspection is also necessary to ensure the consistency of the results when statistical methods are used (Gabriel et al., 2016). Visual inspection has been reported to detect both more (Gabriel et al., 2016; Schoenle and Witzke, 2004) and less (Gabriel et al., 2016; Steppacher et al., 2013) ERP effects than statistical methods.

The subjectivity of visual inspection can be reduced by using criteria concerning, e.g., the effect size or hierarchy of components (Fischer et al., 1999), or by combining the views of 2–3 inspectors (Schoenle and Witzke, 2004; Steppacher et al., 2013). With several raters, full inter-rater agreement is usually required, but the agreement rate is often not reported. This approach resembles the clinical routine used in many hospitals (Gabriel et al., 2016; Kotchoubey, 2015; Steppacher et al., 2013). Results of visual inspection may be confirmed using statistical methods, such as by studying the cross-correlation between different sets of trials (Fischer et al., 1999).

1.2. Average-in-a-time-window ANOVA

ANOVA of voltage amplitude-averages calculated for a time window is the standard of group-level ERP analysis, which is also applicable at single-subject level if inter-trial variation is utilized. ANOVA allows the inclusion of multiple dimensions of data into the analysis as within-subject or between-subjects factors, and avoids the multiple comparisons problem. The amplitude-averaging yields noise resistance and makes the method easily comprehensible. However, if components with opposite deflections are present within suboptimally chosen time window or region of interest, these may cancel each other out as within-trial variation is lost in averaging.

It is possible to utilize data from all electrodes for the analysis of wide-spread effects (Revonsuo et al., 1998), and also interactions between condition and topography may be studied. However, choosing a region of interest (ROI) can increase the sensitivity of the method especially in the case of spatially restricted ERP components. Therefore, the method heavily relies on a priori information of the spatiotemporal location of the ERP component. The choice of ROI should be based on previous information or different dataset and not the data to be analyzed to avoid circularity and implicit multiple comparison problem (Kriegeskorte et al., 2009; Luck, 2014). To introduce more flexibility, and to take into account individual differences in ERP latency, the time window for averaging may also be chosen case-by-case around the peak amplitude observed within the time window of interest (Kappenman and Luck, 2015; Padilla et al., 2006), although this may increase the rate of false positives.

1.3. Cluster-based non-parametric testing

While ANOVA is a parametric approach that uses averages in a time-window, *t*-tests can also be calculated for consecutive time points (Hinterberger et al., 2005), electrodes (Revonsuo and Laine, 1996), or their combination (Bekinschtein et al., 2009). This, however, leads to multiple comparisons problem where *p*-values have to be corrected or the significance threshold needs to be adjusted. Setting an appropriate threshold value is difficult and even the formal approaches have been criticized (Groppe et al., 2011a; Piai et al., 2015). The problems of multiple comparison correction or thresholding may be solved using more universal methods such as permutation tests or cluster-based permutation tests, controlling false discovery rate or generalized false discovery rate (Groppe et al., 2011a).

One of the readily available implementations of cluster-based non-parametric testing is the FieldTrip package for MATLAB (Maris and Oostenveld, 2007). The cluster-mass procedure of FieldTrip includes summing significant *t*-test statistics of adjacent spatiotemporal points to cluster-level statistics (Bullmore et al., 1999; Maris and Oostenveld, 2007). The *p*-value of a cluster is derived from comparison to a Monte Carlo estimate of a permutation distribution, generated by randomly partitioning the trials to different conditions. This makes the test non-parametric, i.e., independent of any assumptions on the distribution of data.

The cluster-based approach leads to weak family-wise error rate control, i.e., a non-specific null hypothesis is being tested (Maris, 2012). Therefore, the test is controlled for multiple comparisons but it provides no information on the significance of individual spatiotemporal points. This leads to high efficiency even with a small number of trials, and increases the sensitivity compared to mass univariate approaches with strong family-wise error rate control (Groppe et al., 2011b; Maris and Oostenveld, 2007). In single-subject analyses, the presence of an ERP component is often of greater importance than the precise electrodes and time points. The methods with weak family-wise error rate control are especially useful with slow cognitive ERPs having a broad scalp distribution.

The cluster-based approach is useful when little a priori information on the topography or latency of an ERP component is available, several almost overlapping components are present, or the topography or latency themselves are of interest (Groppe et al., 2011a). The cluster-based permutation methods such as the one implemented in FieldTrip may be modified by providing additional a priori information concerning the time window, electrodes of interest, or the significance criteria for individual sample points. Since the approach is independent of strictly pre-defined spatiotemporal information, it is well suited to analyses at single-subject level and has been applied in several previous publications (Cruse et al., 2014; Höller et al., 2011; Sculthorpe-Petley et al., 2015).

1.4. Bayesian approach

The voltage-amplitude can also be approximated with regression techniques. Modeling the voltage-amplitude curve using linear model enables taking into account the statistics of the background EEG signal and therefore improves the estimation. It is also possible to model the curve in multiple ways by using different design matrices (Karjalainen, 1997). In this study, the Bayesian approach is adopted as it enables straightforward testing of the presence of ERP.

The idea of Bayesian framework for statistical analyses is based on the Bayes' theorem where the posterior distribution for the parameters conditioned on the observed data can be expressed as the product of the likelihood function that models the observations and the prior distribution for the parameters containing the a priori information (Gelman et al., 1995). The posterior distribution contains all the modeled information of the parameters given the observations, and hence the questions related to the parameters can be answered using basic probability theory. In addition to somewhat simpler framework than frequentist inference, the Bayesian approach is especially powerful if strong a priori information is available. If there is a lack of a priori information, then the prior distributions are more general.

Compared to other analysis methods presented in the current study, the Bayesian regression approach takes into account the user-specified model for the voltage-amplitude curve within the time window of interest. The Bayesian approach is optimal only under the subjective assumptions of the observation models and given the a priori information.

1.5. *t-CWT*

The continuous wavelet transformation combined with Student's *t*-tests (t-CWT) is yet another method for single-subject ERP analysis (Bostanov, 2004; Bostanov, 2015), variants of which have previously been shown to be relatively sensitive and noise-resistant (Bostanov and Kotchoubey, 2006; Daltrozzo et al., 2009; Gabriel et al., 2016; Real et al., 2014; Steppacher et al., 2013). The t-CWT method is designed to extract a set of features from ERP signals and preserve the maximum amount of useful information. The ERP signal is represented in terms of time, amplitude and scale, i.e., the inverse of frequency. These features can then be assessed in a multivariate setting. The data represented as scale-time plots called scalograms can be compared between different conditions by applying *t*-tests at each point and ERP components can be detected from the resulting t-CWT scalograms (Bostanov and Kotchoubey, 2006). The t-CWT method is best suited for analysis of whole ERP waveforms in long time windows, however, it can also be utilized for the study of single ERP components.

The t-CWT method is implemented in a publicly available MATLAB package (Bostanov, 2015). The algorithm uses linear discriminant functions (LDF) to model the different conditions, which can then be used to classify trials into these conditions. Several approaches for performing the comparison between conditions in single-subject analyses have been introduced (Bostanov and Kotchoubey, 2006; Bostanov, 2015; Real et al., 2014). The analysis may either utilize training data from a group of individuals or both the training and testing may be based on a single individual only. If the training and testing phases utilize the same data, the resulting values have been termed "individual biased" as they need to be corrected using randomization tests not implemented in the published package. In the "individual split-half" method, a subset of trials from the participant of interest is used as a training set, and LDFs from the training phase are then used to study the remaining trials. Another method called "individual hold-out" is computationally more demanding because it is based on excluding one trial at a time and using the remaining trials to create LDFs for the classification of the excluded trial. The "group hold-out" approach is analogous with the individual hold-out method but it uses data from several individuals: one person is excluded, the LDFs are formed using data from the remaining persons and the data from the initially excluded person are classified using the LDFs based on the group-level data.

Despite the promising results obtained with the t-CWT method, the optimal approach among those described above and in the literature has not yet been fully established. In addition, not all previous studies have specified the parameter values used in the analyses (Daltrozzo et al., 2009; Steppacher et al., 2013). Due to the lack of a standard approach for t-CWT analysis, we have chosen to include in the present study several approaches which are available in the published software package. However, since the focus of this study lies in methods readily available for use and applying them as implemented in published packages, we did not apply modifications not available in the published software package.

1.6. *N400 effect*

We utilized the N400 ERP component as an example to elucidate how the choice of analysis method influences the detection rate of complex cognitive ERPs. N400 is typically observed in the time window of 250–600 ms post-stimulus for linguistic stimuli and it peaks around 400 ms (Kutas and Hillyard, 1989). The difference between N400 components for congruous and incongruous stimuli is called the N400 effect. N400 effect has been widely used in the clinical setting at single-subject level to test whether semantic processing is present in patients with disorders of consciousness (Balconi and Arangio, 2015; Beukema et al., 2016; Daltrozzo et al., 2009; Erlbeck et al., 2017; Hinterberger et al., 2005; Kotchoubey et al., 2005; Kotchoubey, 2005; Schoenle and Witzke, 2004). Further, its occurrence is associated with good neurological outcome (Rohaut et al., 2015; Steppacher et al., 2013). N400 has broad topography on the scalp but it is strongest in the centroparietal region (Duncan et al., 2009). N400 effect is strengthened by attention (Erlbeck et al., 2014; Holcomb, 1988) and active task (Cruse et al., 2014; Erlbeck et al., 2014). Ideally, N400 effect should be studied using stimuli that are unique throughout the experiment, since its amplitude decreases when the same stimulus is heard repeatedly (Van Petten et al., 1991). Consequently, the number of available stimuli is limited and the need for powerful analysis method is highlighted.

1.7. *Aims of the study*

The incentive for the study is to help both researchers and clinicians to make informed decisions on the selection of single-subject analysis methods and to contextualize and interpret results of ERP studies. In the current study, we investigated the detection rate of visual inspection, ANOVA of average amplitudes in a time window, cluster-based non-parametric testing, Bayesian approach, and t-CWT in single-subject analysis of auditory N400 effect. Each method was used with its conventional and most reasonable parameters, as reported in the previous literature and available in published software packages. Given that the magnitude of N400 effect is dependent on, e.g., active responding to stimuli, the analysis methods were tested using data from both active responding and passive listening paradigms, to simulate the case of unresponsive patients and to produce two conditions with different signal-to-noise ratios. We further examined, quantitatively and qualitatively, which characteristics of the individual ERPs (such as amplitude, timing, topography, length, morphology, and noisiness of signal) were associated with the ability of visual inspection, ANOVA of average amplitudes in a time window, cluster-based non-parametric testing, and Bayesian approach to detect N400 effect.

2. Materials and methods

2.1. *Methods of data collection*

2.1.1. *Participants*

The study was performed at the Turku University Hospital, Finland, after approval of the local Ethics Committee (ClinicalTrials.gov Identifier NCT01889004, Part 1). The participants were 79 healthy 20–30-year-old males (median 23 years). They were right-handed by self-report and had normal hearing determined using Entomed SA 50 screening audiometer (Entomed MedTech AB, Malmö, Sweden). The data reported here are from the screening and baseline experiment of a series of five experiments investigating mechanisms of anesthesia.

Group-level baseline data of 47 participants have previously been reported (Kallionpää et al., 2018). Only males were included because of the radiation exposure related to a subsequent positron emission tomography study. Written informed consent was acquired according to the Declaration of Helsinki.

2.1.2. Stimuli

The stimuli were 310 Finnish auditory high cloze probability sentences prepared along the lines of a previous study (Revonsuo et al., 1998). Specifically, twenty psychology students were asked to fill in the missing last word of the sentences, using a word that first comes to mind and fits the context. Different inflected forms of the same word and clear synonyms were combined to the most common form. Sentences with a resulting cloze probability of at least 50% qualified for the study, and the mean cloze probability of the 310 sentences included in the study was 83.3% (sd 16.2). The sentences were randomly assigned into two equal sized groups. Sentences in the congruous group were kept unchanged, while the last words of sentences in the incongruous group were replaced with context-incompatible words which were matched for lemma frequency, inflection, word class, and number of syllables (Laine and Virtanen, 1999). The first phoneme of the incongruous last word had to differ from its congruous counterpart whenever possible. For example, “Someone knocked on a door” was changed to “Someone knocked on a wife”. The resulting congruous and incongruous sentence groups did not differ in terms of last word lemma frequency, sentence word count and number of syllables in the last word (Mann–Whitney U, $p > 0.05$ for all).

The sentences were digitally recorded by a female native Finnish speaker, and their amplitudes were normalized. A 1 s pause was recorded before the last word of each sentence to avoid the phonetic cues of the last word mixing with the second last word, and the duration of the silence was digitally adjusted to 1000 ms (Ford et al., 1996). The stimuli were divided into a practice block of 10 sentences and two blocks of 150 experimental sentences, each block including 75 congruous and 75 incongruous randomly chosen sentences. There were no significant differences between the two blocks of stimuli in terms of last word lemma frequency, sentence word count and number of syllables in the last word (Mann–Whitney U, $p > 0.05$ to all). The stimuli were presented in the same sequence to all participants.

2.1.3. Equipment and procedure

Data were collected with NeurOne 1.3.1.26 software and Tesla #MRI 2013011 and #MRI 2013012 amplifiers (Mega Electronics Ltd., Finland). The EEG tracing was recorded using a 64-channel EasyCap Active electrode cap that had sintered Ag/AgCl active electrodes placed according to the 10–10 electrode system. EEG was referenced online to FCz and the ground electrode was placed at AFz. Horizontal and vertical eye movements were recorded using four additional electrodes. EEG was recorded with a sampling rate of 1000 Hz with amplifier low-pass filter having half-amplitude threshold of 360 Hz (transition band 250–498 Hz) and high-pass filter of 0.16 Hz (6 dB/octave). The sentence stimuli were presented with Presentation 17.0 stimulus delivery and experimental control software system (Neurobehavioral Systems Inc., CA, USA). All the stimuli and instructions were delivered via headphones.

The subject rested eyes closed on a bed, holding response handles. Stimulus-free baseline was recorded for 2 min. After this, participants were instructed to carefully listen to the stimuli and were told that their memory regarding the sentences would be tested at the end of the experiment. The N400 experiment began with ten practice sentences, after which the actual stimuli were presented in two blocks. Each sentence was followed by a response cue, which was a 100 ms long sine sound played 1 s after the end of the sentence. The response cue was followed by 2.3 s silence before the next sentence started. In the first block, the participants were asked to indicate whether the sentence was congruous or incongruous by squeezing either the right or the left response handle after each stimulus (active paradigm), and in the second block, the task was to carefully listen to the stimuli without responding (passive paradigm). The fixed paradigm order was used in order to obtain baseline measurements for the subsequent anesthesia study reported elsewhere (Kallionpää et al., 2018). The duration of stimulus blocks was 18 min 20 s, and 17 min 53 s, respectively. Handedness for responses was balanced across participants. Another stimulus-free baseline of 2 min was recorded after the stimulus paradigm.

2.2. Data analysis methods

2.2.1. Preprocessing

The preprocessing of the EEG signal was conducted with MATLAB R2013b (MathWorks Inc., USA) and EEGLAB 13_4_4b-toolbox. The signal was downsampled to 250 Hz and re-referenced to the average of the channels TP9 and TP10 (mastoid average). High-pass filtering was performed with non-causal Blackman-windowed sinc-FIR-filter (transition band width 0.2 Hz, passband ripple 0.02% and stopband attenuation −74 dB) using the half-amplitude threshold of 0.1 Hz. The low-pass filtering was performed with a corresponding filter of half-amplitude threshold of 20 Hz and transition bandwidth of 4 Hz.

The trials were segmented −1000–1500 ms relative to the last words of sentences, and epochs containing artifacts were identified by visual inspection with help of independent component analysis (ICA). A median of 6 (range 0–20) trials per participant were removed in the active paradigm, and 4 (range 0–22) trials in the passive paradigm. ICA was run again for the pruned dataset and the components related to eye movements were removed. The noisy channels were interpolated (mean 0.67, median 0, range 0–7 channels per participant). Baseline was corrected using −200–0 ms prestimulus period. In addition to event-related epochs, 75 + 75 epochs of length 2.5 s were randomly segmented from the 2 min stimulus-free baseline period recorded before and after the stimulus blocks and were preprocessed similarly to the event-related epochs. These epochs were used to model the background noise of EEG in the Bayesian regression method. After these preprocessing steps, the data analysis was conducted with five different methods: visual inspection, ANOVA of time-windowed averages, cluster-based nonparametric testing, Bayesian approach, and t-CWT.

2.2.2. Visual inspection

Single-subject averaged ERPs were visually inspected for both active and passive paradigms. Each participant's data were averaged over trials for both paradigms and plotted with Brain Vision Analyzer 2.0 (Brain Products GmbH, Germany) to obtain the average ERP plot for each individual and paradigm. The plots showed the time window −200–1000 ms in 27 channels evenly distributed over the scalp. Congruous and incongruous stimuli were presented as separate curves within the same figure, allowing direct comparison. The figures were presented in randomized order to three raters, who independently evaluated whether N400 effect was present. N400 effect was defined to be present if all of the three raters agreed on it. Agreement between inspectors was evaluated using Fleiss's kappa and within rater pairs using Cohen's kappa.

2.2.3. Average-in-a-time-window ANOVA

The amplitudes of each trial were averaged in the time window 300–600 ms post-stimulus (Kutas and Hillyard, 1980). The analysis was restricted to 13 channels from the centroparietal area (Cz, C1, C2, C3, C4, Pz, P1, P2, P3, P4, CPz, CP1, CP2). All statistics were computed using SPSS Statistics 22 (IBM Corp., NY, USA). Repeated-measures analyses of variance (ANOVAs) were performed at trial-level separately for each participant to compare congruent and incongruent trials and to allow for repeated measures from the 13 electrodes (Kotchoubey et al., 2005). P-values smaller than 0.05 were considered as evidence of the presence of N400 effect. All the comparisons were separately calculated for the active and passive paradigm.

2.2.4. Cluster-based non-parametric testing

Cluster-based non-parametric testing was carried out utilizing the FieldTrip toolbox (Oostenveld et al., 2011) for MATLAB. The congruous and incongruous conditions were compared using one-tailed independent samples *t*-tests in the time window of 200–800 ms post-stimulus (Cruse et al., 2014), and for both paradigms separately. Spatiotemporally adjacent samples were clustered by summing their *t*-values if a significant effect was detected simultaneously in at least two channels having a maximum distance of 40 mm in the standard head model. The *t*-value of the cluster was compared with the *t*-value distribution of 1000 random permutations to obtain a Monte Carlo estimate of cluster *p*-value. An alpha threshold of 0.05 was used, and the clusters with lower *p*-values were defined as indicative of statistically significant N400 effect.

2.2.5. Bayesian method

Linear regression-based estimation was performed based on a novel Bayesian approach, which is described in more detail in a separate technical report (Pesonen et al., 2019). In this approach, we explicitly model the voltage amplitude curve in the time window of 0–800 ms relative to the stimulus, and find the posterior distribution for the parameters that define N400 component for each subject at each of the 13 channels from the centroparietal area. The stimulus-free background EEG signal was assumed to be a stationary Gaussian process with moments evaluated from the stimulus-free epochs recorded before and after the stimulus blocks. The early part (0–296 ms) of the voltage-amplitude after stimulus presentation was assumed equal for the congruent and incongruent stimuli. The latter part (300–800 ms) was assumed to differ by an additive component in the curve under different types of stimuli.

Using these assumptions and a Gaussian prior distribution for the parameters, the Gaussian posterior distribution was evaluated in closed form. The subject-level posterior distribution was evaluated from channel-level posterior distributions, by assuming a common additive N400 component to the investigated channels. If the probability of a negative N400 effect within a time window 300–600 ms was > 95%, N400 effect was concluded to be present.

2.2.6. t-CWT

The t-CWT method was applied using the publicly available MATLAB package t-CWT 2.01 (Bostanov, 2015). The example data provided in the package was used as a template when importing the pre-processed epochs into the analysis. The time window of interest was set to 300–600 ms and the analysis was restricted to the region of interest of 13 channels from the centroparietal area. The cutoff scale S_c was 50 ms corresponding to the cutoff frequency of 20 Hz. We wanted to restrict the analysis to N400 effect which was possible only by limiting the time window and cutoff scale, because restricting the analysis to a negative extremum in a broader time window (like in Bostanov and Kotchoubey, 2006) is not implemented in the package. Log-grid sampling rate R was 15 points per scale which has previously been shown to result in optimal ratio of efficiency and computational time (Bostanov, 2015). The fade-in time T_{in} was 20 ms, fade-out time T_{out} was 200 ms and the eigenvalues that represented 99% of variance in principal component transformation were retained.

The presence of N400 effect was assessed using the methods implemented in the t-CWT 2.01 package: individual split-half, individual hold-out, group hold-out and individual biased. The individual split-half method was utilized in two separate analyses: in the first analysis, the data were split 50%/50% into training and test sets, and in the second analysis, 80% of the data were utilized for training and, similarly to Bostanov (2015), 20% were included in the test set. The a priori error rate was computed based on the actual number of trials in each condition. In the split-half and group hold-out approaches, Hotelling's T^2 test *p*-values smaller than 0.05 were considered as evidence of the N400 effect while for the individual hold-out method the binomial distribution *p*-values for the difference of actual and a priori error rates are presented. Although the individual biased method has previously been reported to be highly efficient (Bostanov and Kotchoubey, 2006), its results are presented without the randomization test correction that would be needed to compensate the accumulation of chance bias as the correction is not implemented in the package.

2.2.7. Performance evaluation of different methods

The differences between the subsets of participants in whom N400 effect was detected by visual inspection, ANOVA, cluster-based non-parametric testing or Bayesian method were evaluated quantitatively by formal comparisons of epoch characteristics and qualitatively by visual classification. To avoid unnecessary complexity and excessive multiple comparisons, the t-CWT method with its alternative testing approaches was not included in quantitative and qualitative comparisons between different methods.

For the quantitative comparisons, N400 effect size and its maximum and median values were calculated from the difference waves for each individual. The standard deviation, kurtosis and skewness were calculated over trials using the time-windowed amplitude averages, and between sample points averaged over trials. Each analysis was restricted to the time window from 300 to 600 ms in the Cz channel. The maximum and average of global field power (GFP) (Lehmann and Skrandies, 1980) were calculated separately for congruent and incongruent trials using average referenced data between 300 and 600 ms. The participants were stratified by whether a significant N400 effect was detected by each method and the resulting subgroups were compared with independent samples *t*-tests separately for each of the four methods in the two paradigms. In addition, the participants in whom N400 effect was not detected with any method were compared to those with a significant effect according to at least one method. As each quantitative feature was tested five times in the two paradigms, all values were compared to the Bonferroni corrected α-level of 0.005.

In the qualitative comparison of the analysis methods, the same N400-average-figures that were used in the context of visual inspection method were visually classified based on timing (typical/early/late/inconsistent), topography (typical/local/frontal/inconsistent), length of effect (typical/short/long/inconsistent), morphology (typical/typical but rough/multiple-peaked/flat/flat but rough/inconsistent), alpha synchronization (not prominent/strong) and overall signal quality (normal/noisy). The classification was performed by one rater in randomized order.

3. Results

The experiment was successfully completed in all participants. In the active paradigm, the participants responded to a median of 100% (range 90–100%) of the 150 stimuli, and 94% of the participants had response rate ≥ 99%. Median of 100% (range 96–100%) of the responses were correct.

3.1. Visual inspection

Visual inspection of N400 effect resulted in substantial agreement between the raters with Fleiss's kappa value of 0.68 and Cohen's kappa values of 0.78 (A–B), 0.68 (B-C) and 0.60 (A–C) for rater pairs. At least two raters reported N400 effect in 118/158 ERP-figures, while all three reported the effect in 105 plots. N400 effect was detected in at least one of the two paradigms for 67 (85%) participants, when agreement of three raters was required (Table 1).

3.2. Average-in-a-time-window ANOVA

The numbers of individual participants with significant ($p < 0.05$) effects are shown in Table 1. N400 effect was detected in at least one of the two paradigms for 54 (68%) participants.

Table 1
The number of participants with detected N400 effect.

	n (% of all 79 participants)					
	Active	Passive	Active & passive	Only active[a]	Only passive[a]	Neither
Visual	56 (70.9%)	49 (62.0%)	38 (48.1%)	18 (22.8%)	11 (13.9%)	12 (15.2%)
ANOVA	43 (54.4%)	33 (41.8%)	22 (27.8%)	21 (26.6%)	11 (13.9%)	25 (31.6%)
Clustered	39 (49.4%)	25 (31.6%)	17 (21.5%)	22 (27.8%)	8 (10.1%)	32 (40.5%)
Bayes	60 (75.9%)	60 (75.9%)	50 (63.3%)	10 (12.7%)	10 (12.7%)	9 (11.4%)
t-CWT						
Split-half 50%/50%	19 (24.1%)	11 (13.9%)	5 (6.3%)	14 (17.7%)	6 (7.6%)	54 (68.4%)
Split-half 80%/20%	13 (16.5%)	7 (8.9%)	3 (3.8%)	10 (12.7%)	4 (5.1%)	62 (78.5%)
Group hold-out	34 (43.0%)	27 (34.2%)	14 (17.7%)	20 (25.3%)	13 (16.5%)	32 (40.5%)

[a] Without cases detected in both active and passive paradigm.

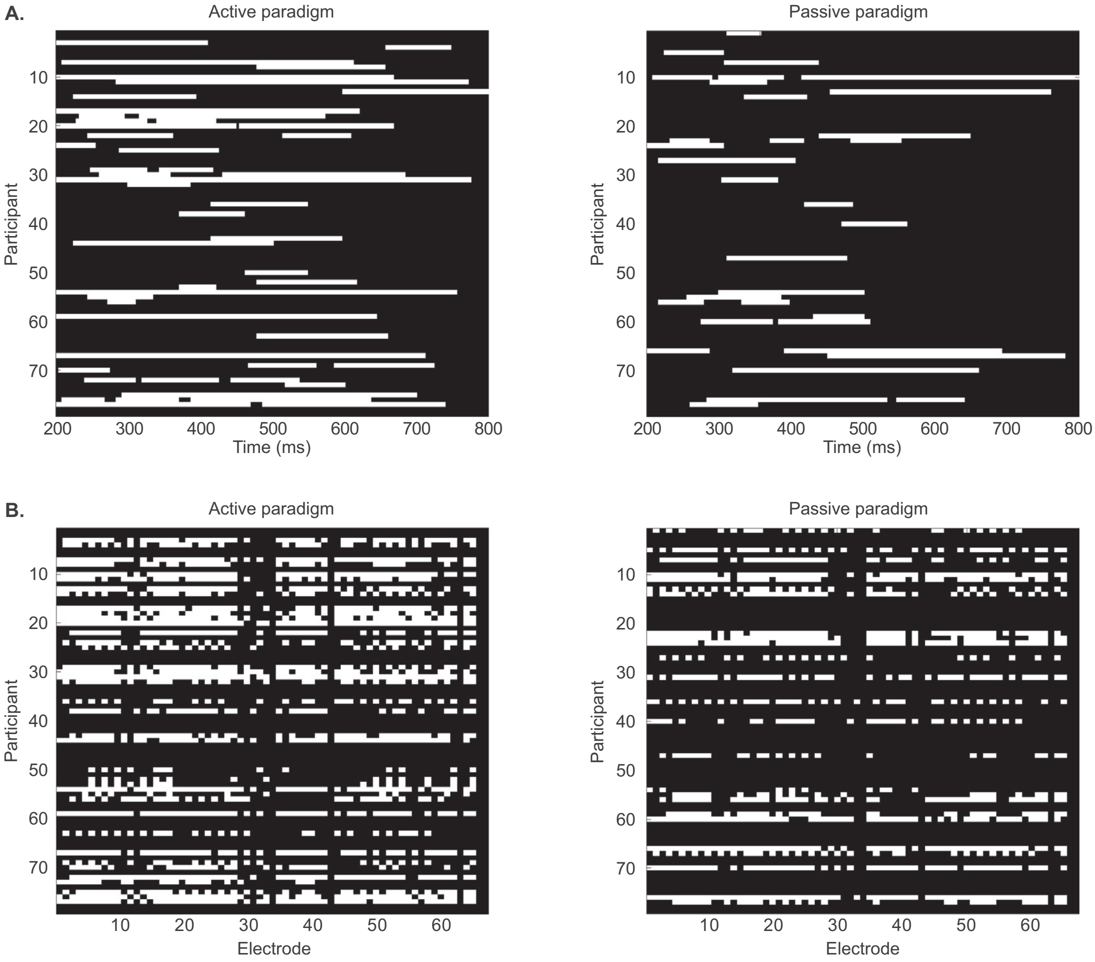

Fig. 1. Clusters that reached statistical significance in terms of time (A.) and electrodes (B.). The sample points and electrodes that belong to the statistically significant clusters are marked with white color.

3.3. Cluster-based non-parametric testing

The effect was observed at single-subject level in 47 (59%) participants in at least one paradigm (Table 1). The significant clusters identified by the algorithm mostly interposed 200 and 600 ms and had a broad scalp distribution (Fig. 1).

3.4. Bayesian method

The computed probabilities for N400 detection for 79 subjects in active and passive setting are illustrated in Fig. 2, and the performance results for 95% detection probability are shown in Table 1. N400 effect was detected in at least one of the two paradigms for 70 (89%) participants.

3.5. t-CWT

The detection results based on the Hotelling's T^2 test *p*-values are shown for the t-CWT split-half method, conducted with two different ratios of trials split in training and test sets, and group hold-out method (Table 1). The full t-CWT analysis results are presented in Supplementary Table 1. The results showed high variation and low concordance between different approaches. The N400 effect was detected in at least one paradigm in 32% (25/79) of participants with 50-%/50%

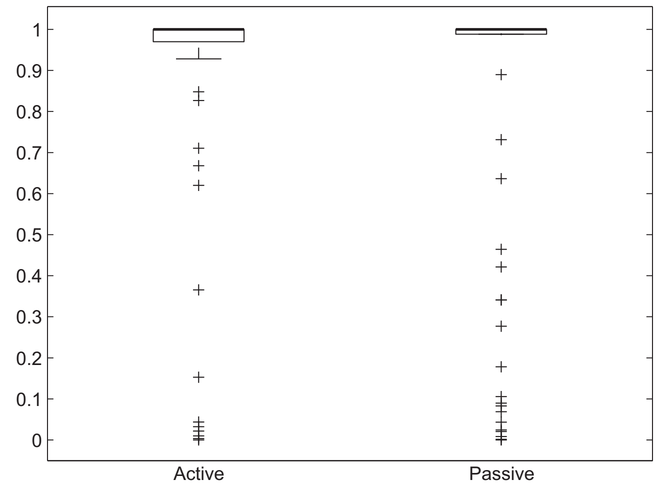

Fig. 2. Tukey boxplot of the posterior probabilities of N400 effect as computed using Bayesian regression. The performance summary of the Bayesian approach can be found in Table 1, when the detection limit is at least 95%. The median was 1.00 in both paradigms (thick bar).

split-half method, 22% (17/79) with 80-%/20% split-half method, and 59% (47/79) with group hold-out method.

3.6. Performance evaluation of different methods

Due to the variety of different t-CWT approaches (Supplementary Table 1), the t-CWT results were omitted from the following performance evaluation of different methods. The individual *p*-values and effect sizes for visual inspection, ANOVA, cluster-based non-parametric testing and Bayesian method are reported in the Supplementary Table 2. N400 effect was detected by at least one of the four methods (visual inspection, ANOVA, cluster-based non-parametric testing and Bayesian method) in 86% (68/79) of participants in the active paradigm and in 86% (68/79) of participants in the passive paradigm (Fig. 3). All four methods detected N400 effect in the same 37% (29/79) of participants in the active paradigm and in 25% (20/79) in the passive paradigm. The second largest intersection between methods was that of visual inspection, ANOVA and Bayesian method. Bayesian regression turned out to be the most lenient method, and resulted in 7 subjects in the active and 17 subjects in the passive paradigm in whom N400 effect was not observed with the other approaches. Also visual inspection detected N400 effect in almost all participants who were also identified with ANOVA and cluster-based methods (Fig. 3).

Visual inspection covered 82% (56/68) and 72% (49/68), ANOVA 63% and 49%, cluster-method 57% and 37%, and Bayesian regression 88% and 88% of the subjects with N400 effect identified by at least one method in the active and passive paradigms, respectively. Out of the participants detected in at least one paradigm, 57% (38/67) were detected in both paradigms with visual inspection, 41% (22/54) with ANOVA, 36% (17/47) with cluster-method and 71% (50/70) with Bayesian regression. In four (5%) participants N400 effect was not identified by any of these four methods in either active or passive paradigm.

We examined which characteristics of the individual ERPs were associated with the ability of visual inspection, ANOVA, cluster-based non-parametric testing and Bayesian method to detect N400 effect. With all methods separately and with the combination of the four methods, the cases with detected N400 effect had more negative amplitude, maximum and median of N400 effect ($p < 0.005$) in Cz electrode compared with the cases without detected effect. The only exception was the maximum of participant-wise average in the passive paradigm where the difference was statistically significant only with ANOVA and the cluster-method. Other measures did not differ significantly between the cases with and without detected effects using any method ($p > 0.005$).

Hence statistical comparisons did not reveal where the differences between visual inspection, ANOVA, cluster-based non-parametric testing and Bayesian method in the detection of N400 lie, the averaged ERPs of each participant were explored qualitatively (Table 2). The cases where N400 effect was not detected by any of the four methods were characterized by inconsistent timing, topography, length or shape. The Bayesian method was the only to detect several cases which were inconsistent in terms of these factors. Although effects typical in terms of timing, topography, length or shape were mainly detected with all the other methods, cluster-method detected only half or less of them.

4. Discussion

There are various methods to analyze ERPs on single-subject level, but it is often difficult to evaluate the effects of methodological choices on the results of a given study. The aim of our study was to investigate, utilizing five different analysis methods, the detection rate of N400 effect on single-subject level in active and passive paradigms. We observed substantial differences between the methods: N400 effect was detected in 16–76% of 79 participants in the active and 9–76% of the participants in the passive paradigm (Table 1). N400 effect was not detected by any method in 8 participants in the active and 10 participants in the passive paradigm, but only two of them were the same subjects. ANOVA was the only method where all positive observations overlapped with at least one other method. The t-CWT and cluster-based technique identified the smallest numbers of cases with N400 effect while the Bayesian regression was the most liberal statistical method in this study. However, none of the used methods detected all

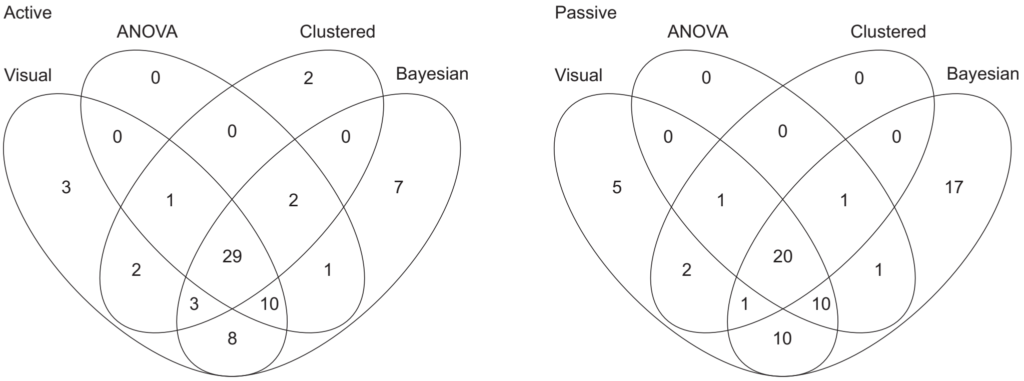

Fig. 3. The numbers of participants in whom N400 effect was detected by visual inspection, ANOVA of average amplitudes in a time window, cluster-based non-parametric testing, and Bayesian approach.

Table 2
N400 effect detection rate of different analysis methods by the qualitative properties of the ERP. Cases from both active and passive paradigms are included.

	Class (n)	Visual (66%)	ANOVA (48%)	Clustered (41%)	Bayes (76%)	None (14%)
	Total (158)					
Timing	Typical (110)	75%	59%	45%	85%	9%
	Early (29)	72%	34%	48%	52%	14%
	Late (3)	0%	0%	0%	100%	0%
	Inconsistent (16)	6%	6%	6%	50%	50%
Topography	Typical (122)	77%	57%	48%	81%	7%
	Local (5)	0%	0%	0%	60%	40%
	Frontal (13)	77%	46%	38%	69%	15%
	Inconsistent (18)	6%	6%	6%	50%	50%
Length	Typical (86)	83%	56%	43%	85%	5%
	Short (23)	30%	0%	17%	48%	30%
	Long (30)	87%	90%	73%	90%	3%
	Inconsistent (19)	5%	5%	5%	47%	53%
Morphology	Typical (72)	86%	63%	56%	82%	7%
	Typical but rough (22)	95%	82%	59%	91%	0%
	Multiple-peaked (23)	39%	26%	30%	70%	13%
	Flat (15)	60%	27%	20%	73%	20%
	Flat but rough (7)	29%	29%	0%	86%	14%
	Inconsistent (19)	11%	5%	5%	42%	53%
Alpha synchronization	Not prominent (106)	75%	49%	44%	75%	13%
	Strong (52)	50%	46%	33%	77%	15%
Overall signal quality	Normal (113)	75%	52%	45%	78%	10%
	Noisy (45)	44%	38%	29%	71%	24%

cases identified by the other methods.

To our knowledge, N400 effect at single-subject-level has been previously studied in healthy participants using auditory sentence stimuli in six studies (Cruse et al., 2014; Daltrozzo et al., 2009; Hinterberger et al., 2005; Kotchoubey, 2005; Rohaut et al., 2015; Sculthorpe-Petley et al., 2015), utilizing five different methods. While cluster-based nonparametric testing detected N400 effect in 49.4% (active paradigm) and 31.6% (passive paradigm) of participants in the present study, the same method found N400 effect only in 17–26% of participants in previous studies (Cruse et al., 2014; Sculthorpe-Petley et al., 2015). This may indicate that the experimental paradigm of the present study was more powerful. The detection rate of visual inspection, ANOVA and Bayesian regression was higher than that of the cluster-based method and similar to what has been observed in previous studies utilizing other methods. The variants of the t-CWT method were the least sensitive in the current study with detection rates in active and passive paradigm ranging from 8.9% to 43% using the group hold-out method and the two variants of individual split-half method. This is in contrast to previous studies where slightly different variants of t-CWT have detected N400 in 60–80% of participants (Daltrozzo et al., 2009; Kotchoubey, 2005). We acknowledge that the t-CWT-based approaches may not show their full potential in our hands, as restricting the analysis to negative extrema only or the randomization tests needed for the interpretation of the individual biased results are not available in the published software package. With methods other than those utilized in the current study, N400 effect has been detected in 40–42% of participants using consecutive *t*-tests (Hinterberger et al., 2005; Rohaut et al., 2015), an approach that resembles the cluster-based testing. Multiple-linear spatial regression approach based on scalp topographies of voltages has shown N400 effect in 42% of participants (Rohaut et al., 2015). A method based on support vector machine classified the averaged ERPs 92% correctly (Sculthorpe-Petley et al., 2015). ANOVA has not been applied at single-subject level in healthy participants but in a non-comparable patient population the N400 effect was detected in 14% of severely brain-damaged but conscious participants (Kotchoubey et al., 2005).

Due to the lack of firm operational criteria for what constitutes an N400 observation, it may be impossible to evaluate the sensitivity of different methods as it is not known whether the effect cannot be detected independently of the analysis method or whether the method is not able to separate the ERP effect from the noise even when the effect is there. Regardless, it is clear that the choice of analysis methods has significant impact on the obtained results. While in the current study we compared different analysis methods for only the detection of N400 effect, large differences between analysis methods have also been reported with mismatch negativity (MMN) (Gabriel et al., 2016) which reflects more automatic cognitive processing of physical characteristics of a sound than N400. Yet, the performance of the methods used in the current study may differ in the detection of other ERPs.

4.1. Factors to consider in single-subject ERP research

Multiple other factors than just the choice of an analysis method may affect the detection rate of ERPs at single-subject level as discussed below.

4.1.1. Individual differences

As previously reported (Cruse et al., 2014; Daltrozzo et al., 2009; Hinterberger et al., 2005; Kotchoubey, 2005; Rohaut et al., 2015; Sculthorpe-Petley et al., 2015) and evident also in the present study, N400 effect cannot be detected in all healthy participants which is a common limitation with cognitive ERPs (Connolly and D'Arcy, 2000). Different ERP amplitudes between individuals may partly result from the differences in individual anatomy, such as the position and orientation of ERP generators and thickness of the skull (reviewed by Luck et al., 2011). The differences may also be due to SNR or individual's concentration and attention. In addition, women have been shown to have larger and earlier N400 effect compared to men (Daltrozzo et al., 2007), which also indicates a limitation of the present study with only male participants.

Nevertheless, single-subject level cognitive ERPs are considered as possible measures of higher-order brain functions and potential tools in diagnostics of, e.g., patients with disorders of consciousness (Bekinschtein et al., 2009; Rohaut et al., 2015; Steppacher et al., 2013). As patients with, for example, brain injury may show atypical topography and latency of specific ERPs, the absence of an ERP does not evidence the absence of cognitive processing typically related to that ERP, either in patients or in healthy individuals.

4.1.2. Characteristics of the experiment

The characteristics of the experiment, e.g., task and attention to the stimuli, whether the experiment is conducted with closed or open eyes, or modality and type of stimuli, may affect the sensitivity of the method to detect the ERP effect. Similarly to the previous observations (Cruse et al., 2014; Erlbeck et al., 2014), all applied methods detected fewer significant N400 effects in the passive compared to the active paradigm. Bayesian regression was the most robust method from the perspective of attention and task as 71% of participants detected in at least one paradigm were the same in both paradigms. With other methods, the switch from active to passive paradigm resulted in lower N400 effect detection rate. The cases with the effect found in the passive but not in the active paradigm might be explained by the presence of the P3 component which is related to response preparation and which temporally overlaps with N400.

The correct motor responses in the active paradigm indicate that the participants comprehended the sentences, and it seems plausible that the processing of meaning has likely been similar in the passive paradigm, although the detection rate was lower. As the active and passive paradigms were performed always in the same order and the duration of the experiment was relatively long, the vigilance of the participants might have been reduced in the passive paradigm. Because the amplitude of N400 effect was smaller in the passive setting, there was more deviation between different analysis methods in detecting the effect.

In this study, conducting the experiment with closed eyes resulted in alpha contamination that was even more prominent in the passive paradigm. Based on the qualitative analysis, the Bayesian approach and ANOVA were the most resistant methods to alpha interference, and might thus be reasonable choices when the data to be analyzed is contaminated with frequencies that cannot be filtered out.

4.1.3. Selection of analysis method

The methods adopted in the current study used a priori information to a different degree. Visual inspection utilized the least amount of information, as only the averaged amplitude curves were analyzed and the deviance between the trials was discarded. In the case of ANOVA, the within-trial information was lost in the time-windowed averaging. Cluster-based method, t-CWT and Bayesian regression utilized the largest amount of data – Bayesian regression even took advantage of the background-EEG.

In terms of quantitative parameters, only the amplitude of N400 effect was associated with the detection of the effect. Quantitative parameters did not explain the differences between results of visual inspection, ANOVA, cluster-based non-parametric testing, and Bayesian approach. In the qualitative comparison, the observed differences between the methods reflected the properties of the methods. The cluster-method and visual inspection were able to detect early effects as well as, or almost as well as, typically timed effects. However, both methods were easily interfered by alpha synchronization and overall poor signal quality. Bayesian method detected even many of the weak and short effects and was tolerant for alpha synchronization, which explains the substantial number of participants where no other method detected effect, especially in the passive setting. Notably, however, the qualitative analysis was performed using the same figures as with the visual inspection method and may thus be biased relative to that method.

4.1.4. Method-specific choices

We are fully aware that in the current study information utilized by the different methods was not identical, such as the selection of the set of time points and channels to be analyzed. However, each method was employed with the parameters typically used in conjunction with the method and no within-method parameter alterations were implemented. These choices necessarily affected the obtained results and rendered the results incompatible for direct comparison.

In case of visual inspection, the full agreement requirement between the three raters made the method strict, although the agreement between raters was good. The results of ANOVA, Bayesian method and t-CWT would have been weaker and irrelevant without using an a priori known region of interest. In t-CWT, the use of pre-defined time window instead of the entire ERP waveform was mandatory because the current t-CWT 2.01 MATLAB package does not have an option for restricting the analysis to negative effects only. On the other hand, flexibility is the greatest advantage of the cluster-method and visual inspection, as they are individually adaptive and no strict preliminary information on the location of the effect is needed. Therefore, restricting the set of channels to be analyzed would have compromised the cluster-based method and visual inspection.

When using methods that require more restricted a priori information, such as specified length of time window or set of channels, the choice has an impact on the detection rate of the effect. In our study, for example, the time window of 300–600 ms used with ANOVA, Bayesian regression and t-CWT might have been shortened to, e.g., 300–500 ms (Van Petten et al., 1991) and the set of centroparietal channels might have been more restricted. In the Bayesian regression, an oversimplified assumption was made that the ERPs to congruous and incongruous stimuli do not differ before the time point 300 ms, although N400 effect can begin earlier especially when combined with phonological mismatch negativity (Connolly and Phillips, 1994). A large variance Gaussian prior was used for the unknown variables mainly for computational convenience.

The results of a total of five variants of the t-CWT methods are shown in Supplementary Table 1, although only three of them allow direct interpretation of Hotelling's T^2 test *p*-values without additional calculations, such as randomization tests. Notably, the split-half method is sensitive to the ratio of the training and test sets, and the variant with only 20% of trials included in the test set showed the lowest detection rate. This is in contrast to a previous study where a test set with 20% of trials showed sufficient discrimination of whole ERP waveforms from passive oddball task (Bostanov, 2015). The group hold-out yielded the highest detection rate. However, this approach utilizes data from a group of individuals for training, which may not always be available for analyses at single-subject level.

Additionally, the choices made in signal preprocessing produce differences in results between studies. Naccache et al. (2016) have, for example, suggested that when complex analysis methods of EEG are used, artifact removal and inspection of raw data have remarkable impact on results. Also other steps of data preprocessing such as filtering and baseline correction affect the results. Kayser and Tenke (2015) have highlighted the importance of the choice of EEG reference and encouraged the use of reference-free surface Laplacian transform. It has been shown that N400 topography is affected by reference selection (Curran et al., 1993). In the current study, the most conventional and recommended reference, linked mastoids, was used (Duncan et al., 2009), and therefore our results should be compared only with the other N400 studies using the same reference.

4.1.5. Future implications

The controversial results yielded by different methods have led many researchers to utilize multiple methods for single-subject analyses to increase reliability (Rohaut et al., 2015; Sculthorpe-Petley et al., 2015; Steppacher et al., 2013). One possible solution to tackle the true detection rate of different analysis methods for various ERPs might be to use simulated datasets with several parameter modifications, as in some previous reports (Groppe et al., 2011b; Real et al., 2014).

In the current study, visual inspection was a sensitive method, which makes it suitable for evaluating the rationality of the results of the statistical methods. This is supported by, e.g., Steppacher et al. (2013) who found visual inspection of N400 effect to have better specificity for recovery from the disorders of consciousness compared with the more sensitive *t*-CWT method (Bostanov, 2004). The methods that identify N400 effect in many subjects in whom the effect is not detected with visual inspection should be utilized with caution. In the present study, Bayesian regression detected N400 effect in 17% (10/60) of participants in active and 32% (19/60) in passive paradigm that were not detected with visual inspection. In the MMN study by Gabriel et al. (2016), *t*-test method yielded 29% (5/17), cross-correlation method 32% (6/19), *t*-CWT 26% (7/27) and multivariate method 27% (4/15) of cases that were not confirmed by visual inspection. In this sense, our implementation of Bayesian regression produced similar results. Relying blindly on a single statistical method may not be sufficient for drawing conclusions on single-subject ERPs, yet visual inspection also has its drawbacks, such as the difficulty of evaluating inter-trial variation.

If the method requires a precise region of interest and time window, it could be beneficial to fix those using group-level averages of the studied population (Rohaut et al., 2015). However, in patient studies, it may be hard to define whether the ERP effect is real if its location, time range and morphology differ from the same ERP effect in healthy participants. The single-subject analyses of event-related potentials involve the continuous balancing between insufficient individuality (bias toward false negatives) and too extensive individuality (bias toward false positives) (Kotchoubey, 2015). If ERPs are used to support clinical decisions, even carefully performed power analysis might not guarantee that the ERP effect will be detected in every participant. In cases where single-subject ERPs are applied in patient populations, the lack of 100% sensitivity even among healthy individuals complicates the interpretation of the results. Clinical decisions cannot be based on negative ERP test results as the negative results do not rule out the presence of the effect or the respective brain function.

4.2. Conclusion

As we show in the current study, different analysis methods provide results that do not completely overlap. Among the methods used in the present study, ANOVA is a feasible choice in the analysis of event-related potentials in single subjects if the spatiotemporal location of the effect is known and false positive results are to be avoided. The Bayesian method has high sensitivity but its results should be confirmed by using some other additional method. Cluster-based non-parametric testing is a viable choice if conservative results are not an issue and some deviation in latency, topography and duration of the effect is assumed. The detection rate of the t-CWT method was much lower in the current study compared to previous reports which implies a need for further studies to explore the applicability of its different variants. Overall, we wish to emphasize that one method alone may not be sufficient to make informed decisions on the presence or absence of an ERP component, and utilizing two or more different types of analysis methods would add the weight of evidence. We hope this study may help to contextualize and interpret findings from other single-subject analysis experiments.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpsycho.2019.06.012.

Acknowledgements

The authors thank Maria Ek and Teemu Laine for their help with electroencephalogram equipment, Milla Karvonen and Otso Saariluoma for assistance with electroencephalogram recordings, and Ville Rajala for the assistance in data preparation. The authors are also grateful to Minna Tallgren for the opportunity to perform this study in the Intensive Care Unit of Turku University Hospital.

Funding

This work was supported by grants from Academy of Finland [grant numbers 266467, 266434] and Jane and Aatos Erkko Foundation.

References

Balconi, M., Arangio, R., 2015. The relationship between coma near coma, disability ratings, and event-related potentials in patients with disorders of consciousness: a semantic association task. Appl. Psychophysiol. Biofeedback 40 (4), 327–337. https://doi.org/10.1007/s10484-015-9304-y.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L., 2009. Neural signature of the conscious processing of auditory regularities. Proc. Natl. Acad. Sci. U. S. A. 106 (5), 1672–1677. https://doi.org/10.1073/pnas.0809667106.

Beukema, S., Gonzalez-Lara, L.E., Finoia, P., Kamau, E., Allanson, J., Chennu, S., ... Cruse, D., 2016. A hierarchy of event-related potential markers of auditory processing in disorders of consciousness. Neuroimage Clin. 12, 359–371. https://doi.org/10.1016/j.nicl.2016.08.003.

Bostanov, V., 2004. BCI competition 2003–data sets Ib and IIb: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. IEEE Trans. Biomed. Eng. 51 (6), 1057–1061. https://doi.org/10.1109/TBME.2004.826702.

Bostanov, V., 2015. Multivariate assessment of event-related potentials with the t-CWT method. BMC Neurosci. 16 (1), 73. https://doi.org/10.1186/s12868-015-0185-z.

Bostanov, V., Kotchoubey, B., 2006. The t-CWT: a new ERP detection and quantification method based on the continuous wavelet transform and Student's t-statistics. Clin. Neurophysiol. 117 (12), 2627–2644. https://doi.org/10.1016/j.clinph.2006.08.012.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M.J., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imaging 18 (1), 32–42. https://doi.org/10.1109/42.750253.

Connolly, J.F., D'Arcy, R.C., 2000. Innovations in neuropsychological assessment using event-related brain potentials. Int. J. Psychophysiol. 37 (1), 31–47. https://doi.org/10.1016/S0167-8760(00)00093-3.

Connolly, J.F., Phillips, N.A., 1994. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. J. Cogn. Neurosci. 6 (3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256.

Cruse, D., Beukema, S., Chennu, S., Malins, J.G., Owen, A.M., McRae, K., 2014. The reliability of the N400 in single subjects: implications for patients with disorders of consciousness. Neuroimage Clin. 4, 788–799. https://doi.org/10.1016/j.nicl.2014.05.001.

Curran, T., Tucker, D.M., Kutas, M., Posner, M.I., 1993. Topography of the N400: brain electrical activity reflecting semantic expectancy. Electroencephalogr. Clin. Neurophysiol. 88 (3), 188–209. https://doi.org/10.1016/0168-5597(93)90004-9.

Daltrozzo, J., Wioland, N., Kotchoubey, B., 2007. Sex differences in two event-related potentials components related to semantic priming. Arch. Sex. Behav. 36 (4), 555–568. https://doi.org/10.1007/s10508-006-9161-0.

Daltrozzo, J., Wioland, N., Mutschler, V., Lutun, P., Calon, B., Meyer, A., ... Kotchoubey, B., 2009. Cortical information processing in coma. Cogn. Behav. Neurol. 22 (1), 53–62. https://doi.org/10.1097/WNN.0b013e318192ccc8.

Duncan, C.C., Barry, R.J., Connolly, J.F., Fischer, C., Michie, P.T., Näätänen, R., ... Van Petten, C., 2009. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clin. Neurophysiol. 120 (11), 1883–1908. https://doi.org/10.1016/j.clinph.2009.07.045.

Erlbeck, H., Kubler, A., Kotchoubey, B., Veser, S., 2014. Task instructions modulate the attentional mode affecting the auditory MMN and the semantic N400. Front. Hum. Neurosci. 8, 654. https://doi.org/10.3389/fnhum.2014.00654.

Erlbeck, H., Real, R.G., Kotchoubey, B., Mattia, D., Bargak, J., Kübler, A., 2017. Basic discriminative and semantic processing in patients in the vegetative and minimally conscious state. Int. J. Psychophysiol. 113, 8–16.

Fischer, C., Morlet, D., Bouchet, P., Luaute, J., Jourdan, C., Salord, F., 1999. Mismatch negativity and late auditory evoked potentials in comatose patients. Clin. Neurophysiol. 110 (9), 1601–1610. https://doi.org/10.1016/S1388-2457(99)00131-5.

Ford, J.M., Woodward, S.H., Sullivan, E.V., Isaacks, B.G., Tinklenberg, J.R., Yesavage, J.A., Roth, W.T., 1996. N400 evidence of abnormal responses to speech in Alzheimer's disease. Electroencephalogr. Clin. Neurophysiol. 99 (3), 235–246. https://doi.org/10.1016/0013-4694(96)95049-X.

Gabriel, D., Muzard, E., Henriques, J., Mignot, C., Pazart, L., Andre-Obadia, N., ... Moulin, T., 2016. Replicability and impact of statistics in the detection of neural responses of consciousness. Brain 139 (Pt 6), e30. https://doi.org/10.1093/brain/aww065.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 1995. Bayesian Data Analysis. Chapman & Hall, London, UK.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011a. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48 (12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011b. Mass univariate analysis of event-related brain potentials/fields II: simulation studies. Psychophysiology 48 (12), 1726–1737. https://doi.org/10.1111/j.1469-8986.2011.01272.x.

Hinterberger, T., Wilhelm, B., Mellinger, J., Kotchoubey, B., Birbaumer, N., 2005. A device for the detection of cognitive brain functions in completely paralyzed or unresponsive patients. IEEE Trans. Biomed. Eng. 52 (2), 211–220. https://doi.org/10.1109/TBME.2004.840190.

Holcomb, P.J., 1988. Automatic and attentional processing: an event-related brain potential analysis of semantic priming. Brain Lang. 35 (1), 66–85. https://doi.org/10.1016/0093-934X(88)90101-0.

Höller, Y., Bergmann, J., Kronbichler, M., Crone, J.S., Schmid, E.V., Golaszewski, S., Ladurner, G., 2011. Preserved oscillatory response but lack of mismatch negativity in patients with disorders of consciousness. Clin. Neurophysiol. 122 (9), 1744–1754. https://doi.org/10.1016/j.clinph.2011.02.009.

Kallionpää, R.E., Scheinin, A., Kallionpää, R.A., Sandman, N., Kallioinen, M., Laitio, R., ... Valli, K., 2018. Spoken words are processed during dexmedetomidine-induced unresponsiveness. Br. J. Anaesth. 121 (1), 270–280. https://doi.org/10.1016/j.bja.2018.04.032.

Kappenman, E.S., Luck, S.J., 2015. Best practices for event-related potential research in clinical populations. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1 (2), 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007.

Karjalainen, P.A., 1997. Regularization and Bayesian Methods for Evoked Potential Estimation. (Doctoral thesis).

Kayser, J., Tenke, C.E., 2015. Issues and considerations for using the scalp surface Laplacian in EEG/ERP research: a tutorial review. Int. J. Psychophysiol. 97 (3), 189–209. https://doi.org/10.1016/j.ijpsycho.2015.04.012.

Kotchoubey, B., 2005. Apallic syndrome is not apallic: is vegetative state vegetative? Neuropsychol. Rehabil. 15 (3–4), 333–356. https://doi.org/10.1080/09602010443000416.

Kotchoubey, B., 2015. Event-related potentials in disorders of consciousness. In: Rossetti, A.O., Laureys, S. (Eds.), Clinical Neurophysiology in Disorders of Consciousness: Brain Function Monitoring in the ICU and Beyond. Springer, Wien, pp. 107–123. https://doi.org/10.1007/978-3-7091-1634-0_9.

Kotchoubey, B., Lang, S., Mezger, G., Schmalohr, D., Schneck, M., Semmler, A., ... Birbaumer, N., 2005. Information processing in severe disorders of consciousness: vegetative state and minimally conscious state. Clin. Neurophysiol. 116 (10), 2441–2453. https://doi.org/10.1016/j.clinph.2005.03.028.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540. https://doi.org/10.1038/nn.2303.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207 (4427), 203–205. https://doi.org/10.1126/science.7350657.

Kutas, M., Hillyard, S.A., 1989. An electrophysiological probe of incidental semantic association. J. Cogn. Neurosci. 1 (1), 38–49. https://doi.org/10.1162/jocn.1989.1.1.38.

Laine, M., Virtanen, P., 1999. Wordmill, Lexical Search Program. Centre for Cognitive Neuroscience, University of Turku, Turku, Finland.

Lang, A., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S., Aaltonen, O., 1995. Practical issues in the clinical application of mismatch negativity. Ear Hear. 16 (1), 118–130.

Lehmann, D., Skrandies, W., 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalogr. Clin. Neurophysiol. 48 (6), 609–621. https://doi.org/10.1016/0013-4694(80)90419-8.

Luck, S.J., 2014. An Introduction to the Event-related Potential Technique, 2nd ed. MIT Press.

Luck, S.J., Gaspelin, N., 2017. How to get statistically significant effects in any ERP experiment (and why you shouldn't). Psychophysiology 54 (1), 114–122. https://doi.org/10.1111/psyp.12639.

Luck, S.J., Mathalon, D.H., O'Donnell, B.F., Hämäläinen, M.S., Spencer, K.M., Javitt, D.C., Uhlhaas, P.J., 2011. A roadmap for the development and validation of event-related potential biomarkers in schizophrenia research. Biol. Psychiatry 70 (1), 28–34. https://doi.org/10.1016/j.biopsych.2010.09.021.

Maris, E., 2012. Statistical testing in electrophysiological studies. Psychophysiology 49 (4), 549–565. https://doi.org/10.1111/j.1469-8986.2011.01320.x.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.

Naccache, L., King, J.R., Sitt, J., Engemann, D., El Karoui, I., Rohaut, B., Dehaene, S., 2015. Neural detection of complex sound sequences or of statistical regularities in the absence of consciousness. Brain 138 (Pt 12), e395. https://doi.org/10.1093/brain/awv190.

Naccache, L., Sitt, J., King, J.R., Rohaut, B., Faugeras, F., Chennu, S., 2016. Reply: replicability and impact of statistics in the detection of neural responses of consciousness. Brain 139 (Pt 6), e31. https://doi.org/10.1093/brain/aww060.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., **2011**. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011, 156869. https://doi.org/10.1155/2011/156869.

Padilla, M.L., Wood, R.A., Hale, L.A., Knight, R.T., 2006. Lapses in a prefrontal-extrastriate preparatory attention network predict mistakes. J. Cogn. Neurosci. 18 (9), 1477–1487. https://doi.org/10.1162/jocn.2006.18.9.1477.

Pesonen, H., Kallionpää, R. E., Scheinin, A., Sandman, N., Laitio, R., Scheinin, H., … Valli, K. (2019). A novel Bayesian linear regression model for analysing event-related potentials: technical report. Retrieved from https://github.com/hpesonen/Bayesian-linear-regression-for-ERPs

Piai, V., Dahlslätt, K., Maris, E., 2015. Statistically comparing EEG/MEG waveforms through successive significant univariate tests: how bad can it be? Psychophysiology 52 (3), 440–443. https://doi.org/10.1111/psyp.12335.

Real, R.G., Kotchoubey, B., Kübler, A., 2014. Studentized continuous wavelet transform (t-CWT) in the analysis of individual ERPs: real and simulated EEG data. Front. Neurosci. 8, 279. https://doi.org/10.3389/fnins.2014.00279.

Revonsuo, A., Laine, M., 1996. Semantic processing without conscious understanding in a global aphasic: evidence from auditory event-related brain potentials. Cortex 32 (1), 29–48. https://doi.org/10.1016/S0010-9452(96)80015-3.

Revonsuo, A., Portin, R., Juottonen, K., Rinne, J.O., 1998. Semantic processing of spoken words in Alzheimer's disease: an electrophysiological study. J. Cogn. Neurosci. 10 (3), 408–420. https://doi.org/10.1162/089892998562726.

Rohaut, B., Faugeras, F., Chausson, N., King, J.R., Karoui, I.E., Cohen, L., Naccache, L., 2015. Probing ERP correlates of verbal semantic processing in patients with impaired consciousness. Neuropsychologia 66, 279–292. https://doi.org/10.1016/j.neuropsychologia.2014.10.014.

Schoenle, P.W., Witzke, W., 2004. How vegetative is the vegetative state? Preserved semantic processing in VS patients – evidence from N 400 event-related potentials. NeuroRehabilitation 19 (4), 329–334.

Sculthorpe-Petley, L., Liu, C., Hajra, S.G., Parvar, H., ... Satel, J., 2015. A rapid event-related potential (ERP) method for point-of-care evaluation of brain function: development of the Halifax consciousness scanner. . J. Neurosci. Methods 245, 64–72. https://doi.org/10.1016/j.jneumeth.2015.02.008.

Steppacher, I., Eickhoff, S., Jordanov, T., Kaps, M., Witzke, W., Kissler, J., 2013. N400 predicts recovery from disorders of consciousness. Ann. Neurol. 73 (5), 594–602. https://doi.org/10.1002/ana.23835.

Tzovara, A., Simonin, A., Oddo, M., Rossetti, A.O., De Lucia, M., 2015. Neural detection of complex sound sequences in the absence of consciousness. Brain 138 (Pt 5), 1160–1166. https://doi.org/10.1093/brain/awv041.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., McIsaac, H., 1991. Fractionating the word repetition effect with event-related potentials. J. Cogn. Neurosci. 3 (2), 131–150. https://doi.org/10.1162/jocn.1991.3.2.131.