Testing the Successive Time Window Approach with Simulated EEG Data

Alexandra Coroiu

University of Twente

Rob van der Lubbe, Stéphanie van den Berg

23 March 2023

Abstract

The multiple testing problem arises when trying to interpret electroencephalogram (EEG) recordings with no a priori assumptions about the timing or location of the expected effect. Noise and data dependencies increase the probability of observing false positives in highly dimensional EEG data. Therefore, the family wise error rate (FWER) needs to be controlled through appropriate statistical methods. In this study, the performance of the successive time window (STW) approach was tested. The specificity, sensitivity, and precision of this approach were assessed for identifying effects on visual event-related potentials (ERPs). The results of this approach were compared to other popular FWER correction methods. The MNE python library was used to create fully synthetic EEG data for a Monte Carlo simulation. Different data parameters were used to define EEG datasets for a between-condition experiment with visual ERPs from 20 simulated subjects. The data were analysed using the STW approach, classic Bonferroni, and a cluster permutation (CP) method. Local and global type I, and II error rates (ERs), and the false discovery rate (FDR) were used to quantify the performance of these methods. The results of this study show that the STW approach leads to a lower local type II ER compared to Bonferroni, but a higher global type I ER (FWER) when compared against both other methods. The STW approach and Bonferroni provide a similar FDR, with better resolution than the CP method. Therefore, the STW approach does not control the FWER as well as the CP and Bonferroni methods, but it can provide more sensitivity and precision. The performance of the STW approach can be highly improved when some a priori assumptions can be made about the location of an expected ERP effect. The critical *p*-value calculation proposed for the STW approach can be improved by adjusting it according to the noise level in rest state baseline EEG data. Future research can build on the methodology of this study to further validate statistical methods aimed at solving the multiple testing problem in ERP studies.

Introduction

Solving the multiple testing problem is important for ensuring proper interpretation of *electroencephalogram* (EEG) recordings in the study of *event-related potential* (ERP). Noise and spatial-temporal dependencies make it difficult to discriminate ERP effects in EEG data when there are no a priori assumptions about the location and timing of an effect (Groppe et al., 2011a). With so many electrodes and so many time points, a high number of type I errors can occur at separate time-space units. These local type I errors increase the (global) *family wise error rate* (FWER) (Luck & Gaspelin, 2017). Various correction methods can be employed to tackle the multiple testing problem and reduce FWER. These methods vary in their theoretical approach and lead to different results depending on the properties of the analysed data (Fields & Kuperberg, 2020; Groppe et al., 2011a). To thoroughly assess the performance of these statistical methods, simulation studies can be used to evaluate the specificity, sensitivity, and precision of results. Monte Carlo simulations allow the testing of statistical methods on a wide range of synthetic EEG data generated based on varying parameters (Groppe et al., 2011b).

The Study of Electrical Brain Signals

EEG is a popular method for recording brain activity. Electrodes are placed at the scalp level to capture electrical signals originating from neural sources within the brain (Baillet et al., 2001). This method has both advantages and disadvantages. On the one hand, EEG recordings result in highly dimensional data: a voltage measurement value is recorded for each unique combination of time and space, at a high frequency across multiple electrodes. This provides a high temporal resolution (in the order of *ms*) as the source electrical signal quickly travels to electrode sites, and it can be captured with high sampling rates. On the other hand, the measurement of electrical signals at the scalp level is highly sensitive to the conductivity of tissue located between the neural sources and electrode sites (Burle et al., 2015). Tissue volume conduction leads to both *spatial smearing*[1] and *temporal*

[1] Spatial smearing is the spreading of electrical signal from one neural source to a wider area at the scalp level. The source signal travels through head tissue which conducts electricity across space. The signal does not directly propagate to the corresponding head surface, but also to proximal areas. This results in multiple electrodes picking up the electrical signal originating from the same underlying neural source. Furthermore, signal from proximal neural sources can blend at the scalp level, resulting in high activation at intermediate electrode sites that do not correspond to the original source locations.

smearing[2] which create dependencies between proximal data points and ultimately reduce the spatial and temporal resolution of EEG data. Furthermore, EEG picks up not only electrical signals coming from the brain, but also other electrical *noise*. When the signal-to-noise ratio is reduced, it becomes difficult to distinguish the true signal from the noise present in EEG recordings.

EEG recordings can be used to study brain activity in ERP research (Luck, 2005). An ERP is the measured electrical brain activity, averaged over trials of the same (class of) stimulus event(s). In an experimental setup, ERPs are obtained by recording the brain response directly resulting from a sensorimotor or cognitive stimulus presented to participants. The source of the resulting ERP is located in the corresponding brain area for the stimulus (di Russo et al., 2002, 2003), while the shape (defined by frequency and amplitude), timing, and duration of an ERP depend on various properties of the stimulus and internal cognitive processes of participants (Key et al., 2005; Rauss et al., 2011; Woodman, 2010). Therefore, ERPs can be used to study properties of cognitive processing within the brain, and to evaluate the difference between groups of participants or between experimental conditions (Baillet et al., 2001). Additionally, ERPs can be used to study brain *lateralization*[3] that arises from peripheral stimulus events (di Russo et al., 2002). Regardless of the experimental setup, ERP research aims to correctly identify whether an ERP effect is present or not in EEG data, and if present, to define the spatial and temporal boundaries of this effect.

The Multiple Testing Problem

The analysis of EEG recordings in ERP research requires the use of statistical methods that can test for ERP effects in the data. When there are a priori assumptions about the location and timing of the expected ERP effect, confirmatory methods can be used to analyse the EEG data. A priori assumptions can highly reduce the size of EEG datasets (e.g., only a few electrodes are selected), and simplify the interpretation of statistical results. If no a priori assumptions can be made, exploratory methods need to be employed (Groppe et al., 2011a). Then, separate local statistical tests are performed for each *data unit* (a unique combination of

[2] Temporal smearing is the spreading of electrical signal from one time point to the next. EEG electrodes can capture the same source signal over several time points across the scalp. Furthermore, signals that are close together in time, can blend together and appear as one continuous signal, with a recorded peak that represents none of the original source signals.

[3] Lateralization refers to the difference in brain activity between the contralateral and ipsilateral hemisphere relative to the stimulus event. ERPs originate from mirrored sources located in both hemispheres. A lateral event leads to different ERP patterns in each hemisphere (e.g., a stimulus on the right will lead to stronger activation in the left hemisphere).

time x space averaged across participants/trials). In the case of exploratory research, the high dimensionality of EEG data introduces the *multiple testing problem.* The multiple testing problem is the statistical problem introduced by performing a high number of simultaneous tests. Even if each test has a low probability (<5%) of resulting in a false positive (type I error), there is a high probability that at least one of these tests will be a false positive. For example, in the case of testing EEG rest state data with 1000 separate local tests, 50 of them can be expected to result in false positives, which would incorrectly suggest that there is an effect in rest state data.

False positives in EEG analysis are caused by noise and data dependencies. All local tests are sensitive to noise, which may lead to significant effects being identified when there are actually none. The presence of local type I errors ultimately affects the global test result (Luck & Gaspelin, 2017). It increases the probability of potentially identifying a global effect in an experimental condition where no activation is expected (e.g., in resting state EEG). Furthermore, when activation is present, local tests at proximal time-space units are highly correlated due to spatial and temporal smearing (Burle et al., 2015). These data dependencies can also negatively impact the global test results (Kim & van de Wiel, 2008). It becomes more difficult to identify the precise boundaries of an existing ERP effect. Different properties of EEG data might increase data dependencies: electrodes and time windows that are closer together pick up correlated signals, and bandpass filtering might increase temporal dependencies (Tanner et al., 2015). These data dependencies can lead to false positives in the proximity of a true ERP signal.

Family Wise Error Rate Control

The multiple testing problem can be solved by controlling FWER. The FWER is the probability of having at least one false positive (one type I error) result in a “family” of simultaneous tests. It represents the probability of concluding that there is a global effect, while there is nothing going on. Various statistical methods have been developed to control the FWER (Groppe et al., 2011a). The performance of different methods varies according to their design (Fields & Kuperberg, 2020; Groppe et al., 2011b). Some methods are only focused on reducing the type I error rate to ensure high *specificity* of results. Specificity represents the probability of correctly dismissing noise and artifacts introduced by data dependencies. Other methods are also focused on diminishing the type II error rate, to ensure higher *sensitivity* as well. Sensitivity represents the probability of identifying effects that are present in the data. There is often a trade-off between specificity and sensitivity. When methods strongly reduce the type I error rate, more false negatives (type II errors) will occur. Low sensitivity leads to underestimating the boundaries of ERP effects (e.g., only high amplitude peaks can be clearly identified). Methods that aim to only control *generalized family wise error rate* (GFWER) can allow more sensitivity (Groppe et al., 2011a). GFWER is the probability of having a number of false positive results in a "family" of tests that is higher than an allowed threshold. Lastly, some methods are simply focused on reducing the *false discovery rate* (FDR), to ensure *precision* in results, instead of controlling the (G)FWER. Precision is the probability that identified effects are true. High precision means that most positive test results are correct. This can ensure that observed effects are not overestimated due to spatial-temporal smearing.

Non-Parametric

The *cluster-based permutation (CP)* method is a non-parametric approach for controlling the FWER (Maris & Oostenveld, 2007). This method requires three steps: 1) select time-space units with *t*-values larger than a desired threshold; 2) create clusters by grouping selected time-space units together based on spatial and temporal adjacency; and 3) test the significance of these clusters based on *p*-values obtained from non-parametric permutation tests. The multiple testing problem is solved by reducing the number of tests performed: one test per cluster, instead of one per time-space unit. This also solves the problem introduced by data dependencies, since correlated time-space units are grouped together within the same cluster. However, the weakness of this method is that the EEG dataset loses its original spatial-temporal resolution, instead making it more suitable only for identifying broadly distributed effects (Fields & Kuperberg, 2020; Groppe et al., 2011a). This weakness is often ignored in research and the CP method is regularly misused by making statements about specific time-space units within clusters (Sassenhagen & Draschkow, 2019).

Parametric

One way to preserve the spatial-temporal resolution of EEG data is to evaluate each time-space unit separately using a parametric statistical test (Fields & Kuperberg, 2020). Although most parametric tests rely on the assumption of independence which does not hold for EEG data, with a high number of time-space units, parametric methods can become more accurate (Clarke & Hall, 2009). Nevertheless, with a high number of local tests performed, appropriate correction methods are needed to balance type I and type II error rates (Luck & Gaspelin, 2017). Various correction methods have been postulated over the years to control the FWER in parametric tests. The classic Bonferroni correction (1) calculates the critical *p*-value for the local level (α') based on the total number of tests performed. However, this method highly reduces the sensitivity of mass univariate testing (Nakagawa, 2004).

A more recent method for controlling the FWER is the *successive time window (STW)* approach (e.g., Talsma et al., 2001). This method requires two steps: 1) test the significance of each time-space unit according to a critical *p*-value for multiple testing correction 2) adjust the significance of each time-space unit according to the successive window criterion. The successive time window criterion entails that a time window is considered significant only if another adjacent one is significant. A new formula (2) was proposed for calculating the corrected critical *p*-value at the local level (van der Lubbe et al., 2014, 2019). The new critical *p*-value calculation accounts for the comparisons performed between successive time windows, which increases the sensitivity of the method.

$$(1)\ \alpha'_{Bonferroni} = \frac{\alpha}{\mathrm{t} * \mathrm{k}}$$

$$(2)\ \alpha'_{Succesive\ Time\ Widnow} = \sqrt{\frac{\alpha}{(\mathrm{t}-1) * \mathrm{k}}}$$

where:
t is the total number of time windows
k is the total number of electrodes

Research Goal

Although the STW approach has been used before, no formal assessment has been done for its validity as a FWER control method. The goal of this research is to evaluate the validity of the STW approach in an EEG experimental setup with conditions defined by the presentation of peripheral (left/right visual) stimuli. Validity is defined by the performance of this approach in terms of specificity, sensitivity, and precision. An optimal FWER control method should minimize type I errors from rest state data, and it should offer high sensitivity and precision when an ERP effect is present. The performance of the STW approach is investigated through several research subgoals within this experimental setup.

The first five subgoals are defined for assessing the performance of this approach for testing different effects in visual ERPs. *Subgoal I:* Test the performance of the STW approach for identifying the absence of signal in EEG data from a (resting state) baseline condition. *Subgoal II:* Test the performance of this approach for identifying the presence of signal in EEG data from conditions when a peripheral stimulus is present in the right or left visual field. *Subgoal III:* Test the performance of this approach for identifying the signal difference between the two peripheral (left/right) stimulus presentation conditions. *Subgoal IV:* Test the performance of this approach for identifying the absence of brain lateralization in the (resting state) baseline condition. *Subgoal V:* Test the performance of this approach for identifying the presence of brain lateralization from peripherally presented stimuli, as it was used in the original study (van der Lubbe et al., 2019). Additionally, *subgoal VI:* The Bonferroni and the CP method are used as benchmark for the performance of the STW approach for all tests used to evaluate subgoals I-V.

Furthermore, two more subgoals are defined for exploring the performance of the STW approach in different EEG datasets. *Subgoal VII:* Explore the influence of different data parameters on the performance of the STW approach. Data parameters are introduced to determine the signal-to-noise ratio (amplitude, noise level), data dependencies (bandpass filtering, electrode density, time window size) and the presence or absence of a priori assumptions (location, time interval). It is expected that lower performance is associated with lower signal-to-noise ratio, more data dependencies, and a lack of a priori assumptions. *Subgoal VIII:* Explore the influence of the critical *p*-value calculation on the performance of the STW approach across the different data sizes (as resulted from the use of different data parameters). Lastly, *subgoal IX:* Make suggestions for the critical *p*-value calculation to maximize the performance of the STW approach.

Simulation Studies

With the advance of modelling technology, EEG data can now be artificially generated and used for simulation studies. Simulation studies can thoroughly verify the methods that are commonly used in EEG data analysis (Fields & Kuperberg, 2020; Groppe et al., 2011b; Luck & Gaspelin, 2017). The simulation study by Sassenhagen & Draschkow (2019) shows how simulations can be used to highlight pitfalls in the use of current methods. The advantage of using simulated EEG data lays in the fact that the location and timing of a simulated source signal is known. This simulated signal can be used to check the sensitivity and precision of various statistical methods for identifying localized effects. Similarly, simulated rest-state baseline EEG data is certain to be empty of any activation that might otherwise arise due to internal cognitive processes in participants (Luck, 2005). This baseline data can be used to check the FWER control and specificity offered by the statistical methods used for multiple testing correction. Therefore, simulated EEG data can be used to assess the performance of statistical methods and compare results between them.

Furthermore, simulation studies also provide the opportunity for testing statistical methods under different data parameters. Realistic EEG datasets can be created from scratch with the help of anatomical brain and head models, simulated neural signals, *forward modelling*[4] techniques, and artificially generated noise and artefacts (Barzegaran et al., 2019). This allows the creation of Monte Carlo simulations that can cover varying parameter values and create a wide range of realistic EEG data. For assessing statistical methods used in EEG data analysis, the aggregated results across various simulated datasets can provide a more realistic estimate of the methods' performance, compared to simulations that only cover one dataset. For example, Monte Carlo simulations have been used to simulate ERP effects of different shapes (Groppe et al., 2011b), to introduce variance among subjects (Luck & Gaspelin, 2017), or to test methods under varying a priori assumptions (Fields & Kuperberg, 2020). Additionally, Monte Carlo simulations have been used to observe EEG data dependencies under different processing parameters (Burle et al., 2015), or to determine the spatial accuracy of EEG (Liu et al., 2002). Therefore, simulated data can be created to cover a wide range of data parameters which affect the results of EEG analysis.

[4] Forward modelling is the process of calculating the spread of electrical signal from a neural source to the scalp surface. Calculations account for the conductivity of head tissue, resulting in realistic spatial and temporal smearing of the source signal.

Methods

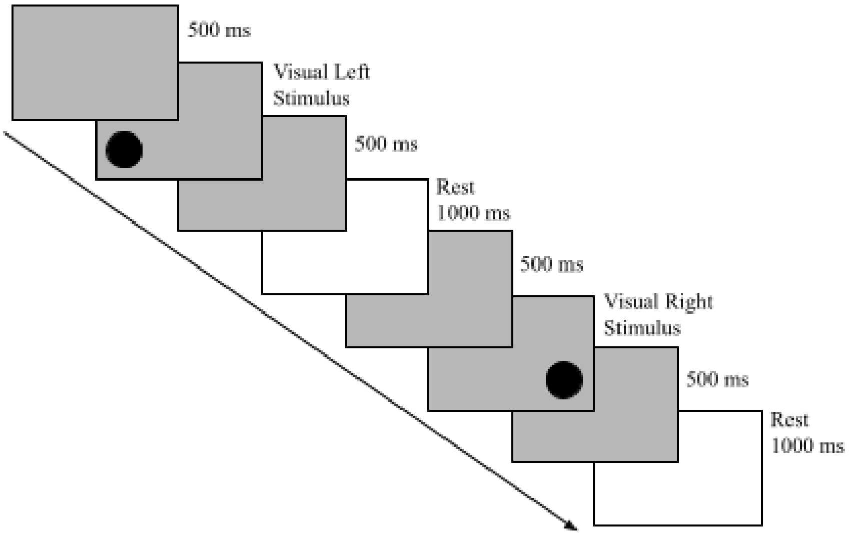
Monte Carlo simulation was used to evaluate the performance of the STW approach. All steps of the project were implemented in Python using libraries for the analysis of neurophysiological data. The MNE library was used for simulating, processing, visualizing, and analysing the EEG data (Gramfort et al., 2013). Data science libraries were used for the advanced scientific computations applied to the data and for visualizing results. The source code is publicly available on GitHub, with instructions on how to run the simulation, analyse the EEG data and evaluate the results (Coroiu, 2022).

Simulation

Realistic EEG datasets were fully simulated for a within-participant experiment with conditions of peripherally presented visual stimuli. The simulation contains three conditions: 1) a baseline – no stimulus presented, 2) visual left – stimulus presented in the left visual field, and 3) visual right – stimulus presented in the right visual field. Data were simulated for 20 participants, with 20 trials for each visual stimulus condition separated by the baseline resting state condition (see Figure 1).

Figure 1

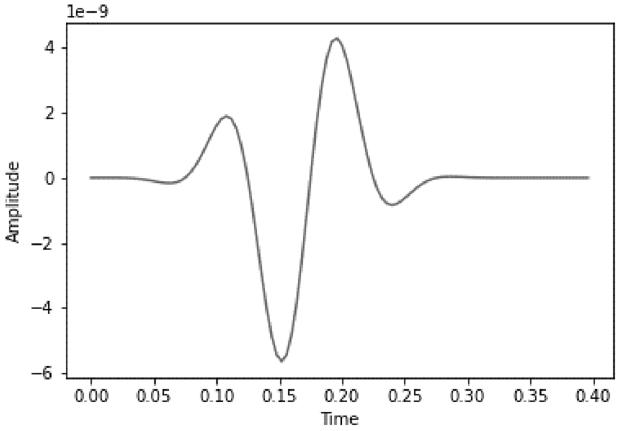
Trials Visual Stimulus Presentation Conditions



Note: One (left/right) visual stimulus trial contains the timeframe (-500ms, +500ms) around stimulus presentation. The baseline trial corresponds to the resting state in between the visual stimulus trials. Each trial amounts to a total of 1000ms.

The neural activation for the visual stimulus conditions was simulated based on a pre-defined function that generates synthetic signals. The signal takes the form of a wave with amplitude at the scale of microvolts (μV). The shape of the wave is constructed through the multiplication of two composite waves (5). The first composite wave (3) is a sinusoid with frequency 10 *Hz*, corresponding to the alpha frequency band. The second composite wave (4) is a gaussian function (gf) which defines the latency and duration of the signal. The gaussian wave peaks at the desired latency and the area between $\pm\ 2.5\sigma$ corresponds to a duration of 200 ms. The resulting synthetic signal resembles the early components of an ERP associated with the presentation of a stimulus in the lower visual field (di Russo et al., 2002, 2003; Rauss et al., 2011). The simulated ERP contains three main peaks (50 ms apart) associated with the visual P1 (C1), N1, and P2 (Key et al., 2005) (see Figure 2). No ERP was introduced in the baseline condition. This corresponds to signal of constant amplitude 0.

Figure 2
Simulated ERP Wave



$$(3)\ sinusoid = \sin(2\pi * 10)$$

$$(4)\ gf \sim N(latency, 0.04)$$

$$(5)\ wave = sinusoid * gf *\ amplitude$$

Note: The timepoint at 0 ms corresponds to the moment of stimulus onset. Visual ERP component P1 peaks at 100 ms, N1 at 150 ms and P2 at 200 ms. This ERP wave was generated with latency = 175 ms and amplitude = 60 µV.

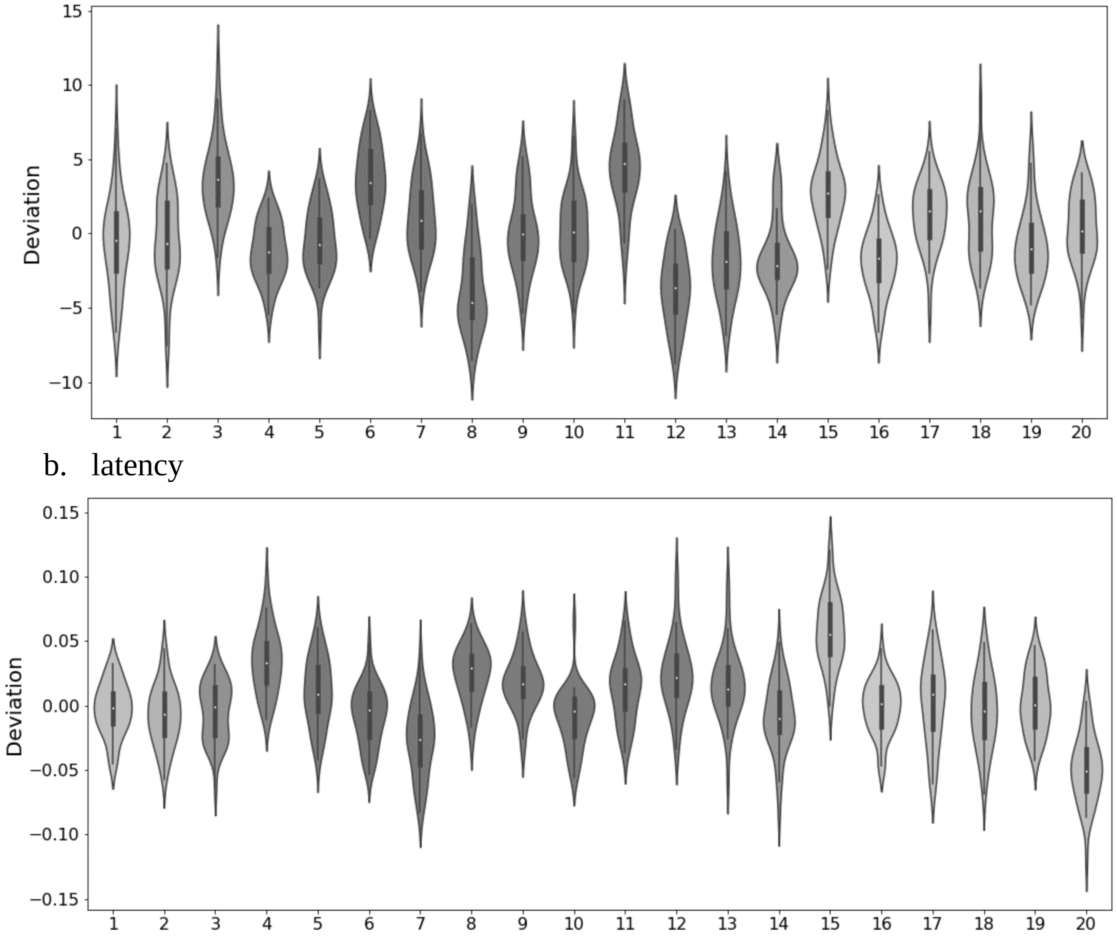
Two simulated ERPs were introduced in all visual stimulus trials (one for a contralateral and one for an ipsilateral neural source). The simulated ERPs vary according to two data parameters: amplitude and latency. These parameters were used with the goal of creating realistic experimental data. Firstly, experimental setup and the properties of presented stimuli affect attentional processes in participants, and lead to varying ERP amplitude levels (Key et al., 2005). Therefore, six different amplitude (μV) pairs were used with different values and ratios between contra and ipsilateral activation: (40,20), (60,30), (60,20), (80,40), (80,30), (80,20). A different dataset was generated for each amplitude pair (Subgoal VII). For the visual left stimulus condition, the activity is stronger in the right hemisphere (contralateral) and weaker in the left hemisphere (ipsilateral), and vice versa for the visual right stimulus condition (di Russo et al., 2002, 2003; Rauss et al., 2011). The activity in the baseline condition is absent, corresponding to an amplitude of 0 μV. Secondly, the latency of the contralateral activity wave was set to 175 ms for all datasets (as in Figure 2). The ipsilateral activity has an additional interhemispheric transfer time (IHTT) delay of 15 ms (di Russo et al., 2002, 2003). Therefore, the simulated ERP wave for the ipsilateral source is defined by a gaussian function with peak at 190 ms.

Furthermore, differences in cognition influence both amplitude and latency (Key et al., 2005; Takemura et al., 2020; Woodman, 2010). Therefore, variability was introduced for the simulated signal at both the participant and the trial level (see Figure 3a,b). Each participant's mean difference from the standard latency and amplitude values is drawn from a normal distribution ($M_{latency} = 0,\ SD_{latency} = 25\ ms$, and $M_{amplitude} = 0,\ SD_{amplitude} = 2.5\ \mu V$), and each trial is drawn from a normal distribution around the participant mean ($M_{latency} = M_{participant},\ SD_{latency} = 25\ ms$, and $M_{amplitude}\ = M_{participant},\ SD_{ampllitude} = 2.5\ \mu V$). The participant and trial variability were pre-generated, and the same values were used across all simulated datasets.

The anatomy of simulated participants was based on the default scan of one sample subject provided in MNE. Therefore, all participants present the same default source space, connectivity, forward model, and conductivity values. First, the simulated activity was generated from the left/right superior occipital gyrus, corresponding to a stimulus peripherally presented in the lower visual field (di Russo et al., 2003). The specific locations of the activity sources in each hemisphere were defined using the default MNE source space data. Second, the default MNE average transformation matrix was used to define the propagation of the signal across the brain. Then, the forward model was generated using the Boundary Element Method (BEM) (Hallez et al., 2007). A standard BEM model was created using the sample subject anatomy and MNE default volume conduction coefficients ($c_{brain} = 0.3,\ c_{skull} = 0.006,\ c_{scalp} = 0.3$). Through the application of the forward model, data was generated for the whole head surface using the initial source activity (Figure 4a).

Figure 3

Participant and Trial Level Variability



a. amplitude

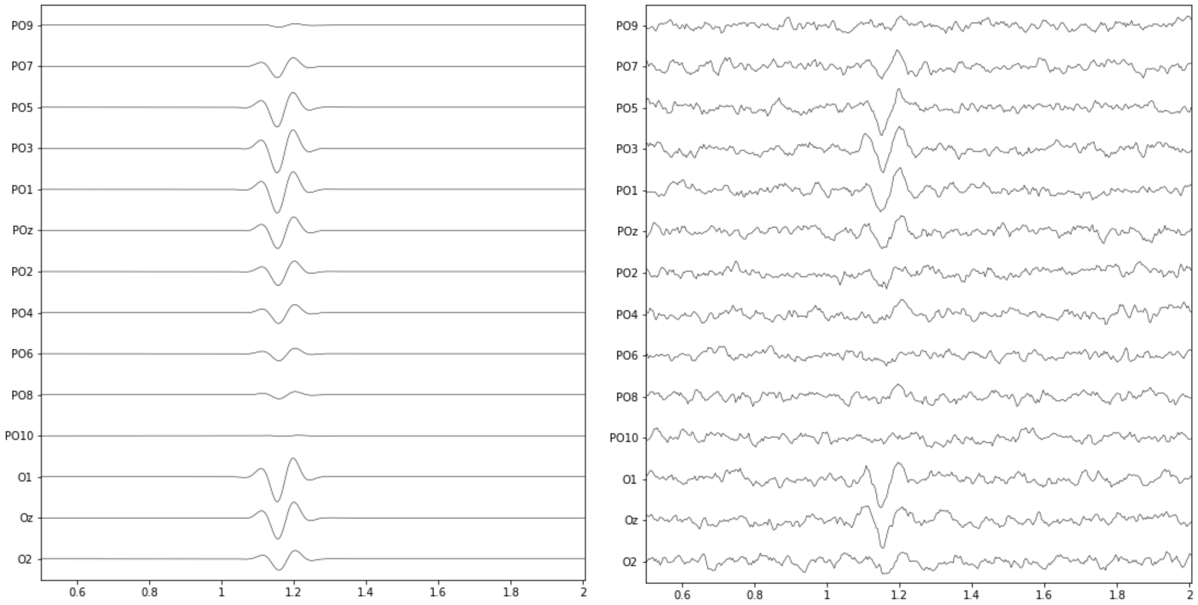
b. latency

Note: The deviation distribution for each participant (numbered 1-20) is composed of the individual trial level differences from the standard amplitude and standard latency. Amplitude deviation is measured in μV, while latency deviation is measured in ms.

The final simulated scalp level signal aims to mimic realistic raw EEG electrode recordings. To achieve this, noise was added on top of the data (see Figure 4b). First, random noise was generated using an ad hoc covariance matrix. The same matrix has been used across all simulated datasets. Then, an infinite impulse response (IIR) filter was used to tailor the generated noise levels into two parameter values (defined by the linear filter coefficients): high noise ($coefficients = [0.1, -0.1, 0.02]$) and low noise ($coefficients = [0.2, -0.2, 0.04]$) (Bekthi et al., 2022). A different dataset was generated for each noise value. Together with the amplitude parameter, the noise parameter determines the signal-to-noise ratios across the simulated datasets (Subgoal VII). Finally, for each participant data was simulated for 86 electrodes, placed according to the extended 10/20 international system (Klem et al., 1999). The recordings were simulated for a sampling frequency of 250 Hz, resulting in time windows of 4ms.

Figure 4

Simulated EEG Data at the Electrode Level



a. before adding noise

b. after adding noise

Note: The values for each electrode represent the signal that is present at corresponding scalp locations after the application of the forward model. The signal for each electrode is measured on an amplitude scale of (± 20μV). The presented time frame (0.5 to 2s) displays the first trial (visual right presentation condition) from the raw EEG data of one participant. The signal for this trial was generated with an amplitude of (60,20) + 0.209μV, and a latency of 175 (after the stimulus presentation event at 1s).

Processing

The raw data of each participant was processed and prepared for statistical analysis (see Table A1 for example). First, the raw data was epoched: the relevant time intervals around stimulus presentation were selected (-500 ms before to 500 ms after stimulus onset) and baseline (-500ms to stimulus onset) correction was applied (Tanner et al., 2016). The application of a band pass filter is another data parameter, as this kind of filtering can introduce artifacts within EEG data due to increased temporal dependencies (Tanner et al., 2015) (Subgoal VII). Datasets were generated with and without the suggested band pass filter (0.1-30 *Hz*). Second, the epoched data was summarized into evoked data. For each participant, trial level data points were averaged together to form the time-space units that will be used for the statistical analysis. Trial data were averaged separately for each of the three experimental condition (baseline, visual right, visual left) (Subgoals I-II).

Then, the data was prepared separately for two types of analysis. *For identifying the difference in signal between the visual stimulus presentation conditions*, the difference between the evoked ERPs from the visual right (vr) and visual left (vl) conditions is computed at each time-space unit (6). Together, the calculated differences become a fourth experimental condition to be tested (Subgoal III). *For identifying brain lateralization*, the difference between hemisphere activation was calculated per electrode pair (e.g., PO7-PO8). The value at the ipsilateral electrode was subtracted from the contralateral electrode for each time point (7). For calculating lateralization of the baseline (Subgoal IV), the right hemisphere was always considered contralateral in this condition. The lateralization values calculated for the visual right and visual left conditions (Subgoal V) were averaged into one visual stimuli lateralization value for each time-space unit (a unique combination of *time x electrode pair*) (8) (van der Lubbe et al., 2019). Since only matching electrode pairs are used instead of separate electrodes, the data prepared for lateralization contains less than half of the total number of electrodes in the initial data (midline electrodes are also excluded).

$$(6)\ difference = \mathrm{u}_{\mathrm{vr}} - \mathrm{u}_{\mathrm{vl}}$$

$$(7)\ lateralization = \mathrm{u}_{\mathrm{contra}} - \mathrm{u}_{\mathrm{ipsi}}$$

$$(8)\ \overline{lateralization} = \frac{lateralization_{vr} + lateralization_{\mathrm{vl}}}{2}$$

Furthermore, different datasets are prepared for analysis based on additional data parameters. The sampling rate (time window size) and electrode density determine the number of separate time-space units to be tested, which ultimately affects the global test performance (Groppe et al., 2011a) (Subgoal VII). Time points are selected for analysis depending on the sampling window size: 4 ms, 12 ms, or 20 ms. Electrodes are selected based on the montage density: 32, 64, or 86. Higher sampling rate and higher electrode density can lead to an increase in data dependencies as the proximity of separate time-space units increases. Two more data parameters are defined by the presence or absence of *a priori* assumptions about the location and time interval of the expected effect (Subgoal VII). These a priori assumptions affect the global test performance as well (Fields & Kuperberg, 2020; Groppe et al., 2011a). For the location assumption, either all electrodes or only the ones associated with the visual field (parietal to occipital electrodes) are selected. For the time interval assumption, either the whole post stimulus time period is selected (0 – 500 ms) or a constrained time frame where a visual ERP is usually expected (50 – 300 ms) (Key et al., 2005).

Analysis

Successive Time Window

Mass univariate parametric testing was implemented using one-sample t-tests. A two-tailed t-test ($df = 19$) was performed at every time-space unit. For the three experimental conditions (baseline, visual right, visual left), the value at each time-space unit was tested against zero to check for the presence (or lack) of localized signal. When comparing the visual right and visual left conditions, the local test against zero checks whether there is a significant difference between the two conditions at each time-space unit (time x electrode). When testing for lateralization, the local test against zero checks whether there is a significant effect between hemisphere activation at each time-space unit (time x electrode pair).

For implementing the STW approach, the multiple testing was corrected using the critical p-value and then the successive window technique was applied (van der Lubbe et al., 2019). First, the STW critical p-value was calculated based on the desired global significance (α = 0.05) and the number of unique time-space units that were tested in each dataset: $(time\ windows - 1) * \ electrodes$, or $(time\ windows - 1) * \ electrode\ pairs$ (for lateralization). The number of time-space units depends on the data parameters used to prepare each dataset (time window size, electrode density and a priori assumptions). The resulting p-value from each local test was compared against the calculated critical p-value. If the p-value of a local test was lower than the critical p-value, then the effect was considered significant for that time-space unit. Afterwards, the local test results were adjusted based on adjacent time windows: a local test was considered significant only if the previous or the next test was also significant (see Table A2 for example).

Bonferroni

Mass univariate parametric testing with the Bonferroni correction was implemented similarly to the STW approach. One-sample two-tailed t-tests ($df = 19$) were performed to check for significant ERP effects at each time-space unit. The Bonferroni critical p-value was calculated based on the desired global significance (α = 0.05) and the number of unique time-space units that were tested in each dataset. If the p-value of a local test was lower than the critical p-value, then the effect was considered significant for that time-space unit.

Cluster Permutation

The CP method (Maris & Oostenveld, 2007) was implemented using the default MNE functions. First, a parametric two-tailed *t*-test (α = 0.05) was performed for each time-space unit. Secondly, selected time-space units were used to create clusters based on the spatial and temporal dependencies of the data. The spatial dependency of the data was modelled by generating an adjacency matrix for the electrodes used in each dataset. The number of generated clusters varies across datasets, depending on the data parameters used (time window size, electrode density and a priori assumptions). Thirdly, cluster-level statistics were calculated by summing the *t*-statistics of all time-space units in a cluster. *P*-values for each cluster were calculated under the permutation distribution of the largest cluster-level statistic among all created clusters. Finally, each cluster was tested for significance (α = 0.05) (see Table A3 for example).

Evaluation

Performance Metrics

The performance of the statistical methods across the different tests (Subgoals I-V) was assessed both at the local and global level (Luck & Gaspelin, 2017). The local level measures whether the local test significance results match the expected effect at each specific time-space unit. The global level measures whether the global test significance matches the expected effect in each test condition. The local and global performance of statistical methods was assessed in terms of specificity, sensitivity, and precision. These metrics were evaluated as *error rates* (ERs) (Luck & Gaspelin, 2017), extracted from a confusion matrix (see Table 1). The type I ER (9) was calculated based on the rate of false positives in all negatives and represents the complement of specificity, while the type II ER (10) was calculated based on the rate of false negatives in all positives and represents the complement of sensitivity. Each confusion matrix was created according to the total number of tests, the expected effect from the simulation and the results of the statistical methods (see Table A5 for example).

Table 1
Confusion Matrix

Tested	Effect	
	Present	Not Present
Significant	TP	FP
Not Significant	FN	TN

Note: TP = true positives, FP = false positives, FN = false negatives, TN = true negatives

$$(9)\ Type\ I\ error\ rate\ = \frac{FP}{FP\ +\ TN} = 1 - specificity$$

$$(10)\ Type\ II\ error\ rate\ = \frac{FN}{FN\ +\ TP} = 1 - sensitivity$$

The local ERs were calculated for each dataset separately (see Table A4 for example). A local time-space unit was expected to be significant if its location and time matched the simulated ERP. The location corresponds to the visual area (parietal to occipital electrodes) and the time interval corresponds to the area between ±1.875σ of the gaussian function used to generate the ERP wave (100 - 250 ms). This time interval contains signal that is above the introduced noise level (ERP components N1, P2), and should therefore test significant. For the local level evaluation, statistics (mean and standard deviation) were used to summarize ERs across datasets. The mean local type I ER represents the probability for a time-space unit with no effect to test as a false positive, while the mean local type II ER represent the probability for a time-space unit with a true present effect to test as a false negative. No local evaluation can be performed for the CP method, as it is not possible to make statements about the significance of specific data units within a significant cluster (Sassenhagen & Draschkow, 2019).

The global ERs are calculated across datasets. With global ERs, the performance of the STW approach can be compared not only against the parametric Bonferroni correction but also against the non-parametric CP method (Subgoal VI) (Groppe et al., 2011b). Global significance for each dataset was assessed based on the local test results. A test was considered globally significant if at least one local test/cluster was found significant (Luck & Gaspelin, 2017). Because of the successive time window criterion that was previously applied for the STW approach, a globally significant test always contains *at least two* successive local time-space units that overstepped the critical *p*-value. On the one hand, a global test was expected to be significant in conditions where an effect was indeed present (in the visual left/right conditions, the difference, and the lateralization calculated from the two visual stimulus conditions). The global type II ER represents the probability for an experiment to result in a false negative for these conditions. On the other hand, no effect is present in baseline conditions, therefore the global tests for these conditions are not expected to be significant. The global type I ER represents the probability for an experiment to result in a false positive for a baseline condition where no effect is present. The global type I ER on baseline conditions quantifies how well the methods control for the FWER.

Furthermore, for the tests where a global effect is identified, statements about the boundaries of the effect(s) can be made based on the significant local tests. However, as not all significant local tests correspond to true effects, erroneous statements can be made. These mistakes are quantified by the FDR (11). FDR is also used to compare the performance of the STW approach against the other methods (Subgoal VI). For the STW approach and Bonferroni, locally significant tests can be used to indicate at exactly what time-space units an effect is present. A significant time-space unit was considered a false positive if its time and location do not correspond to the expected signal. For the CP method, the significant clusters can be used to indicate a broader spatial-temporal range where an effect can be found. A significant cluster was considered a false positive if it contained no time-space units that matched the time and location of the expected ERP effect. Therefore, FDR represents the probability for a significant local test to contain no true effects. The lower the FDR, the more precise statements can be made about the boundaries of a significant global effect.

$$(11)\ FDR\ = \frac{FP}{FP + TP} = 1 - precision$$

Data Parameters

The effect of data parameters (independent variables) was evaluated for the performance of the STW approach (dependent variable) (Subgoal VII). All data parameters used to simulate the different datasets were included in this explorative analysis: amplitude, noise, band pass filtering, window size, a priori time interval, electrode density, and a priori electrode location. The performance of the STW was quantified by the global and local metrics (type I, II ERs and FDR). Data parameters were considered discrete variables, while the performance metrics are continuous. Therefore, performance metrics were compared between the different discrete values of each data parameter.

Critical P-Value Calculation

The critical *p*-value is calculated based on the number of local tests that need to be performed. The number of local tests varies across datasets based on the selected parameters (window size, electrode density, a priori location and a priori time interval). Ideally the critical *p*-value would correct the local significance of tests in such a way that the STW approach would perform optimally for any dataset. Therefore, there should be minimal correlation between the dataset size (independent variable) and the performance of the STW approach (dependent variables). The relation between these two variables was explored through plots and regression analysis (Subgoal VIII). The performance of the STW was quantified by the local and global metrics (type I, II ERs and FDR). Both dataset size and the performance metrics were considered continuous variables.

Results

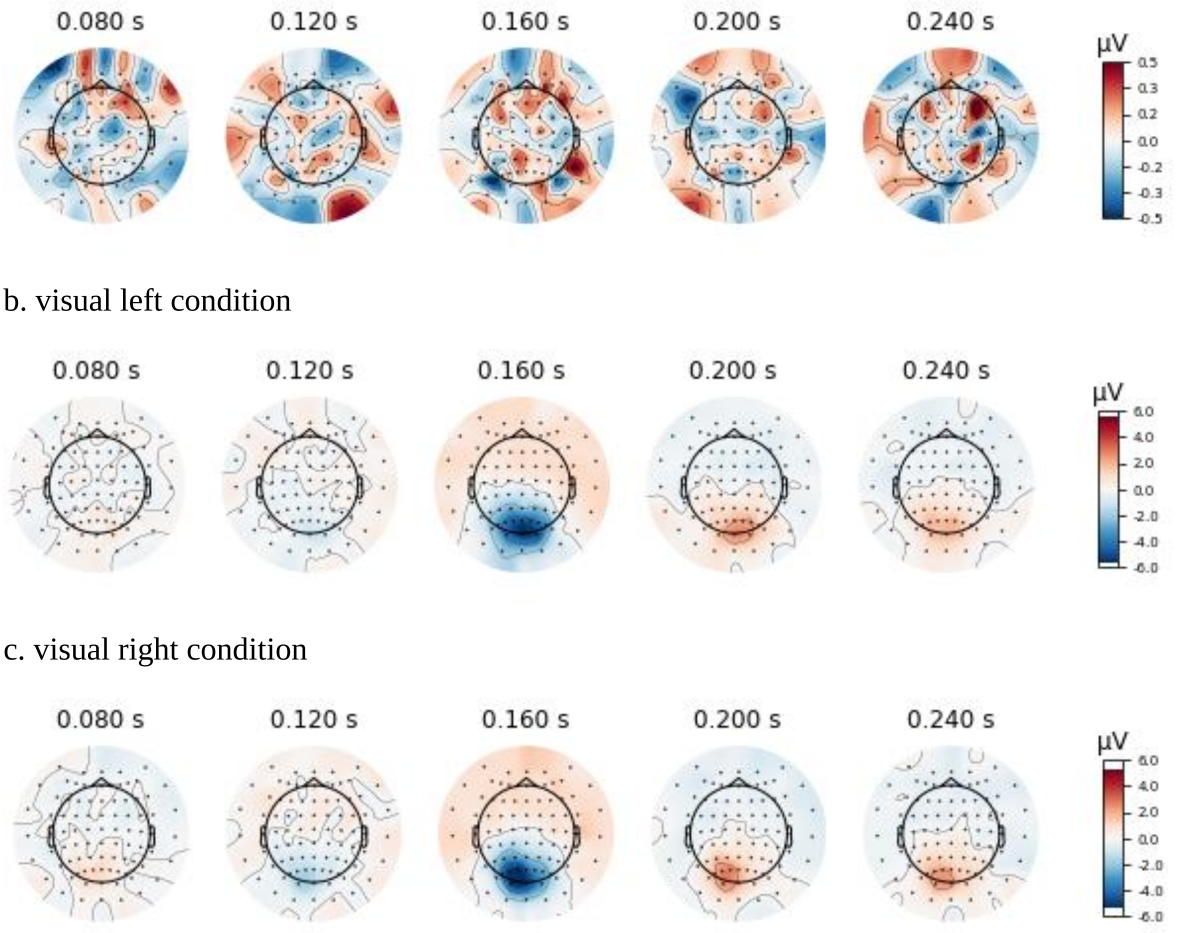
The resulting ERP signal can be observed in topographic maps and graphs generated from the simulated EEG data. The results of the analysis and evaluation of simulated EEG data are summarized to assess the performance of the different statistical methods across experimental conditions at the local and global level (subgoals I-VI). The impact of data parameters and the critical *p*-value calculation was evaluated for the STW approach (subgoals VII-IX).

Simulation

The EEG simulation contains a total of 864 datasets, defined by seven different data parameters (*6 amplitudes values x 2 noise values x 2 bandpass values x 3 window size values x 2 time interval values x 3 density values x 2 location values*). Each dataset contains data for the 20 simulated participants under the three defined experimental conditions. The evoked ERP signal for one participant, averaged across 20 trials for each condition, can be seen in Appendix B. The baseline condition presents no significant ERP activity, with amplitude values staying very low ($\pm\ 0.5\ \mu V$), only displaying a random pattern stemming from the added noise (see Figure 5a). The visual stimulus presentation conditions show negative and positive voltage changes ($\pm\ 6\ \mu V$) around the visual area (parietal to occipital electrodes) at 160 ms and 200 ms (post stimulus onset), which match the simulated ERP signal (see Figure 5b,c). However, the location of the activity across the visual area does not show strong lateralization between the left and right hemisphere. These aspects of the simulation are controlled by the sample subject anatomy (source space, connectivity, forward model and/or conductivity). The location and orientation of the ipsilateral and contralateral source lead to the signal propagating towards the centre of the visual area (parietal to occipital electrodes). The activation for the visual left condition is especially centred around the midline electrodes. Furthermore, given spatial and temporal smearing the two ipsilateral and contralateral sources are likely to blend together into one observable signal.

Figure 5

Grand Average Topographic Maps


a. baseline condition

b. visual left condition

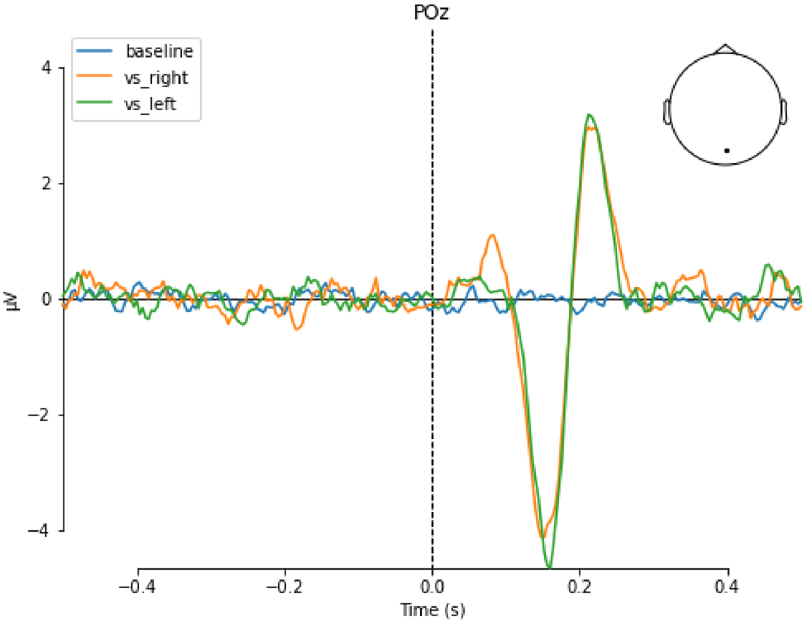
c. visual right condition

Note: Values for each electrode are averaged over trials. This data was generated under the following values for data parameters: (60,20) amplitude, high noise, and no band pass filter applied. Negativity is plotted downwards.

The signal at the visual POz electrode captures the presence of the simulated ERP wave for the visual left and visual right conditions (see Figure 6). The two main ERP peaks (N1, P2) are clearly visible around 175 ms after right/left visual stimulus (vs) presentation for both conditions. The smaller P1 component is not as clearly distinguishable from noise (especially for the visual left stimulus condition). The signal from the baseline condition displays only noise level data. The noise levels observed in the final simulated EEG recordings fall within estimations of noise picked up by EEG equipment (0.3 to 2 μV) (Teplan, 2002).

Figure 6

Grand Average ERP at the POz electrode



Note: Values for each time point are averaged over trials. This data was generated under the following values for data parameters: (60,20) amplitude, high noise, no band pass filter applied. Negativity is plotted downwards.

Local Type I/II Error Rates

The performance of the STW approach at the local level shows that the method sacrifices a lot of sensitivity when testing separate time-space units (see Table 2). On the one hand, the type I ER is low for all conditions ($M < 10\%$,), especially for the baseline, difference, and lateralization conditions ($M < 1\%$,). The slightly higher type I ER for the visual stimulus conditions can be an indication that spatial smearing leads to signal being spread to time-space units where it is not expected. On the other hand, type II ER is high ($M \cong 56\%$) for the visual stimulus presentation conditions, and very high when testing for lateralization in visual stimulus conditions ($M = 79.58\%$) and for signal difference between visual stimulus conditions ($M = 98.68\%$). This shows that the STW approach misses most of the expected effects at local time-space units. For comparison, almost all local effects are missed regardless of condition ($M > 87\%$) with the Bonferroni correction.

Table 2
Mean Local Error Rates Across Datasets

condition	metric	Bonferroni	STW
baseline	type I ER	0.0001	0.0021
	type II ER	-	-
visual left	type I ER	0.0062	0.046
	type II ER	0.8761	0.5591
visual right	type I ER	0.0065	0.0726
	type II ER	0.8942	0.5673
difference	type I ER	0.0002	0.0034
	type II ER	0.9999	0.9864
lateralization baseline	type I ER	0	0.0031
	type II ER	-	-
lateralization visual stimuli	type I ER	0.0006	0.0084
	type II ER	0.9584	0.7941

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Global Type I/II Error Rates

Comparing the performance of the STW approach and other FWER correction methods at the global level shows that they can lead to different results depending on the tested condition (see Table 3). All methods correctly detect the (left/right) visual stimulus conditions in all datasets (Type II ER = 0%). When testing the lateralization from visual stimulus conditions, the global type II ER is very low for the STW and CP method (ER < 2%), but higher for the Bonferroni correction (ER = 14%). The CP and Bonferroni method fail to correctly identify the difference between the right and left visual conditions in most datasets (Type II ER ≅ 90%). While the Bonferroni method simply has very low sensitivity, the results for the CP method can be explained by the lack of visible lateralization in the simulated data. Since both the visual left and visual right conditions result in activation closer to the centre of the visual area (parietal to occipital electrodes), there is no broadly distributed difference between the two signal patterns. This makes it difficult for the CP method to detect the effect. As the STW approach is more sensitive to localized effects, this method can more easily identify localized differences at the time-space unit level (Type II ER = 21.64%).

Nevertheless, for the STW approach, the sensitivity to localized effects leads to the baseline conditions incorrectly testing significant at the global level in about half the datasets

(Type I ER baseline = 51.5%, Type I ER lateralization baseline = 45.25%,). This shows that there is a random chance for at least one local type I error. Meanwhile, the Bonferroni and CP method test correctly for both baseline conditions at the global level (Type I ERs < 5%). This shows that generally, when no global effect is expected, the CP method does not produce any significant cluster. Therefore, CP and SWT can both identify a global effect under conditions where a general effect is expected, but the STW approach is too sensitive to noise when testing baseline conditions at the global level. This shows that the STW approach does not provide strict control for the FWER. The Bonferroni method provides the desired FWER but sacrifices sensitivity.

Table 3
Global Error Rates

condition	metric	Bonferroni	STW	CP
baseline	type I ER	0.0162	0.515	0.0347
	type II ER	-	-	-
visual left	type I ER	-	-	-
	type II ER	0	0	0
visual right	type I ER	-	-	-
	type II ER	0	0	0
difference	type I ER	-	-	-
	type II ER	0.897	0.2164	0.9144
lateralization baseline	type I ER	0.0104	0.4525	0.0281*
	type II ER	-	-	-
lateralization visual stimuli	type I ER	-	-	-
	type II ER	0.14	0.0116	0.0162

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).
**Value excludes ten datasets for which no clusters could be formed.*

False Discovery Rate

The rate of false positives in globally significant tests depends on the tested condition (see Table 4). For the tests where a global effect has been identified in the visual left/right conditions, most units/clusters contain true signal (FDR ≅ 25%). Similarly, for the tests where a global lateralization effect has been identified in the visual stimulus conditions, almost all units/clusters contain true lateralization (FDR < 10%). However, for the difference between conditions, more than half of the significant clusters/units are false positives for the CP

(66.22%) and Bonferroni (72.47%) methods. The STW approach performs only slightly better, with 43% FDR in significant time-space units. Therefore, all methods can be used to make fairly precise statements about the boundaries of an effect in visual stimulus conditions and lateralization, but neither can be used to make precise statements about the difference between visual stimulus conditions.

Table 4
Mean FDR Across Globally Significant Datasets

condition	Bonferroni	STW	CP*
baseline	-	-	-
visual left	0.2093	0.2762	0.2206
visual right	0.2209	0.3285	0.233
difference	0.7247	0.437	0.6622
lat. baseline	-	-	-
lat. visual stimuli	0.0543	0.0933	0.0695

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0). **For the CP method, the values represent the cluster FDR, compared to Bonferroni and STW, where the values represent the data unit FDR*

The Impact of Data Parameters

Tables for the evaluation of each parameter can be found in the Appendix C.

Amplitude

The amplitude value and ratio between contra and ipsilateral activation affects the performance of the STW approach. At the local level, datasets with lower amplitude values and ratios have lower type I ER and higher type II ER when testing for the presence of signal and lateralization. At the global level, type II ER is almost 30% higher for the smallest amplitude datasets: (40,20), (60,30), when testing for the difference between visual stimulus conditions. Furthermore, even when differences are identified at the global level, the FDR for low amplitude datasets is up to 15% higher. Nevertheless, these datasets also have lower FDR for identifying the presence signal in visual stimulus conditions. The lower type I ER and FDR could be explained by a possible reduction in spatial smearing associated with a lower amplitude source signal. Overall, smaller amplitude values and ratios increase the probability of missing effects when these expected effects are small but can also allow more precise results when expected effects are larger. Therefore, the STW approach offers less sensitivity, but more precision for low amplitude data.

Noise

The level of noise also affects the performance of the STW approach. At the local level, datasets with higher noise have higher type II ER when testing for the presence of signal and lateralization. This could be explained by the application of baseline correction: lower parts of the true signal get eliminated if average noise levels from the baseline are close to signal levels. Then, only the signal corresponding to the tip of ERP peaks (components N1, P2) remain after correction. Interestingly, global type I ER is higher for baseline signal and lower for baseline lateralization in high noise datasets. This could be explained by the subtraction done for the lateralization between the two hemispheres, which most of the time will cancel out high noise values. Furthermore, high noise data leads to lower FDR when testing for the presence of signal in visual stimulus conditions. Overall, these results show that it is more difficult to identify signal in noisy EEG data, but when signal is identified, the results might be more precise (potentially corresponding only to signal peak amplitudes). Therefore, the STW approach offers less sensitivity, but more precision for high noise data.

Band Pass

The application of the band pass filter does not seem to clearly affect the performance of the STW approach. No differences can be observed at the local level, either in type I ER, type II ER or FDR. However, the datasets where a bandpass filter was applied have a 15% higher global type I ER on the baseline condition. This suggests that the bandpass filter leads to some small local artefacts being introduced, which increase the probability of "at least one significant" time-space unit in each dataset. However, these artifacts are too small to impact the performance of the STW approach when signal is present in the data. Since there are no differences in conditions where an effect is expected, the dependencies between time points introduced by temporal smearing do not seem to impact the performance of STW method for this experimental setup.

Time Window Size

The window size affects the performance of the STW approach. While no clear differences can be identified at the local level, the window size greatly affects global performance. Global type I ERs on the baseline conditions increases for smaller time windows by two (12ms) to four times (4ms). The FDR for difference and lateralization also increases for smaller time windows by up to 20%. This shows that small time windows increase the number of local type I errors per dataset. However, the largest time window (20ms) can lead to almost 50% higher global type II ER when testing for the difference between the visual stimulus presentation conditions. These results show that a larger window size can result in true effects being missed. This might be a specific problem for the short ERP signal simulated in this study. With peaks that are only 50 ms apart, a window size of 20 ms makes it likely that time points with peak amplitudes are excluded from analysis. Additionally, the successive time window comparison used to adjust local significance might easily label peak expected signal as not significant because far apart adjacent time windows do not include high enough amplitudes. Therefore, smaller time windows decrease specificity and precision, but time windows that are too large (relative to signal frequency) decrease the sensitivity of the STW approach.

Time interval

The a priori assumption for the time interval also affects the performance of the STW approach. At the local level, type I ERs are higher when testing for the presence of signal in the visual stimulus conditions when a priori time assumptions are made. This can be explained by the reduction in the total number of time-space units, which leads to a higher critical *p*-value for local tests. However, global type I ERs stay lower for the baseline conditions if time assumptions are made. Furthermore, when testing the difference between visual stimulus conditions, FDR is also 17% lower. By exclusively testing a time interval where an effect is truly present, the ratio between data containing true signal and data potentially containing type I errors increases, which improves FDR. Therefore, making assumptions about the time interval of an effect can increase the specificity and precision of results for the STW approach.

Electrode Density

Electrode density affects the performance of the STW approach. Higher electrode density leads to higher type II ER at the local level when testing for signal and lateralization in visual stimulus conditions. This can be explained by the increase in the total number of time-space units, which leads to a lower critical *p*-value for local tests. At the global level, the lowest density datasets (31 electrodes) have up to 12% less type I ER on the baseline conditions. This shows that using less electrodes decreases the probability of “at least one local type I error” to occur in baseline conditions. Therefore, the STW approach offers more specificity and more sensitivity for low electrode density data.

Electrode Locations

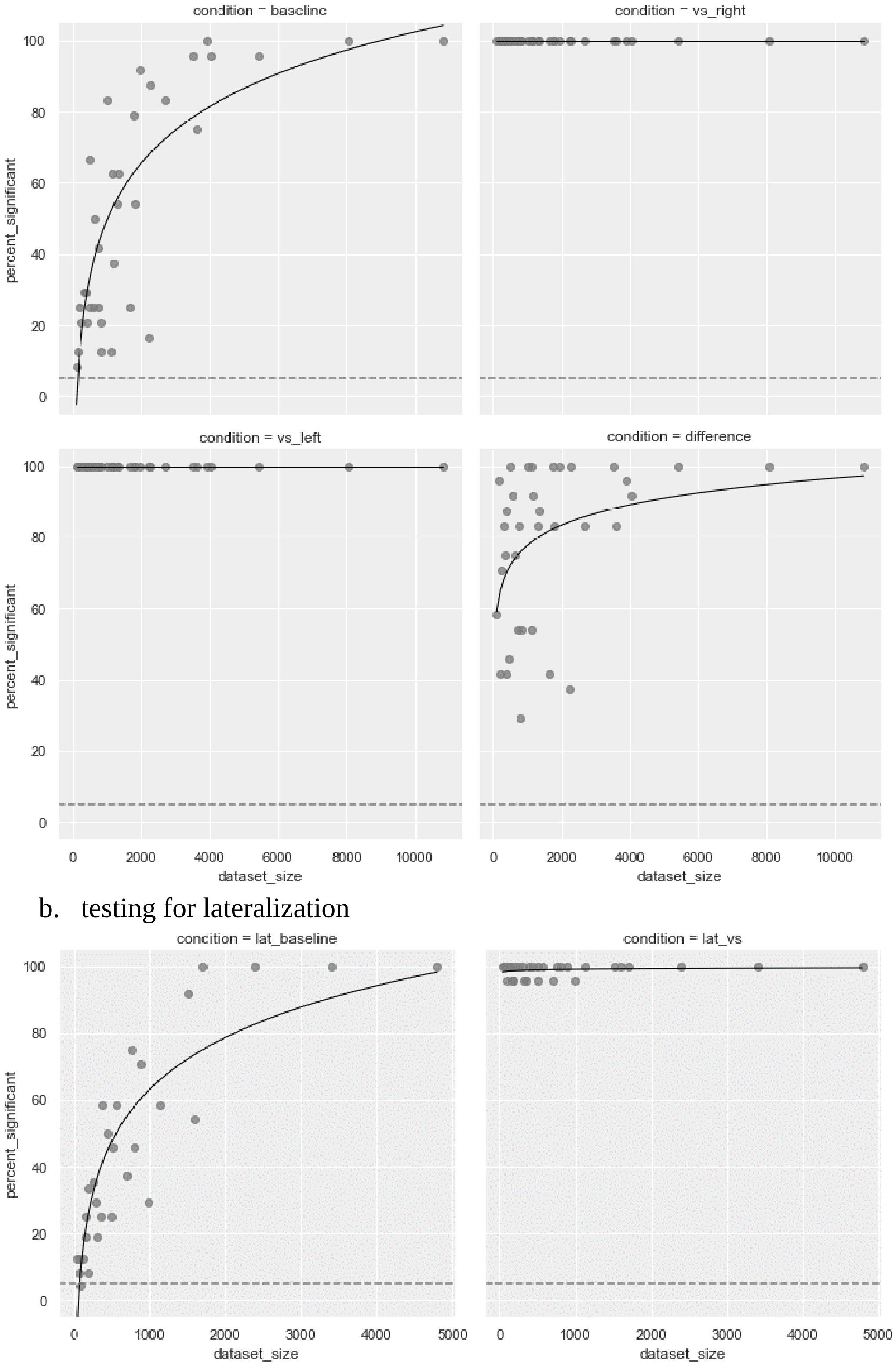
A priori selection of the electrode location highly impacts the performance of the STW approach. At the local level, both higher type II ERs and higher type I ERs are observed if no location assumptions are made when testing for signal in the visual stimulus presentation conditions. Up to 26% higher type I ER can also be observed at the global level for the baseline conditions and FDR is up to 50% higher if no location assumptions are made. By exclusively testing an area where an effect is truly present, the amount of type I errors coming from other locations is reduced. On the one hand, this helps to reduce the impact of noise and spatial smearing on results, but on the other hand it makes it impossible to identify related effects in other brain areas if they exist (e.g., pre-frontal activation from attentional processes). Therefore, a priori location assumptions increase the specificity, sensitivity, and precision of results for the STW results.

Critical *P*-Value Calculation

Dataset size is correlated with the performance of the STW approach. At the global level, there is lower probability for small datasets to test significant at the global level on baseline conditions (see Figure 7a,b). Datasets that contain less than 2000 local tests (baseline) or 1000 local tests (lateralization baseline), have a probability lower than 50% to test as a global false positive. This probability quickly increases to almost 100% for datasets with 4000 (baseline) or 2000 (lateralization baseline) local tests. As expected, there is better control for the FWER with smaller datasets. However, smaller datasets (<2000) lead to an increased probability of global false negatives when testing for very small difference effects. The presence of signal and lateralization can be correctly identified regardless of dataset size. Furthermore, dataset size is also positively correlated with FDR (see Figure D1). Datasets with less than 2000 local tests (baseline), have higher precision (FDR < 40%) for significant global tests on conditions where an effect is present. At the local level, there is no clear correlation between data size and the local type I/II ER (see Figures D2-3). The higher critical *p*-value for very small datasets ($\lesssim 1000$ local tests) seems to decrease local type II ER (by up to 10%), while local type I ER is mostly unaffected.

Figure 7

Global Significance Across Datasets



a. testing for signal

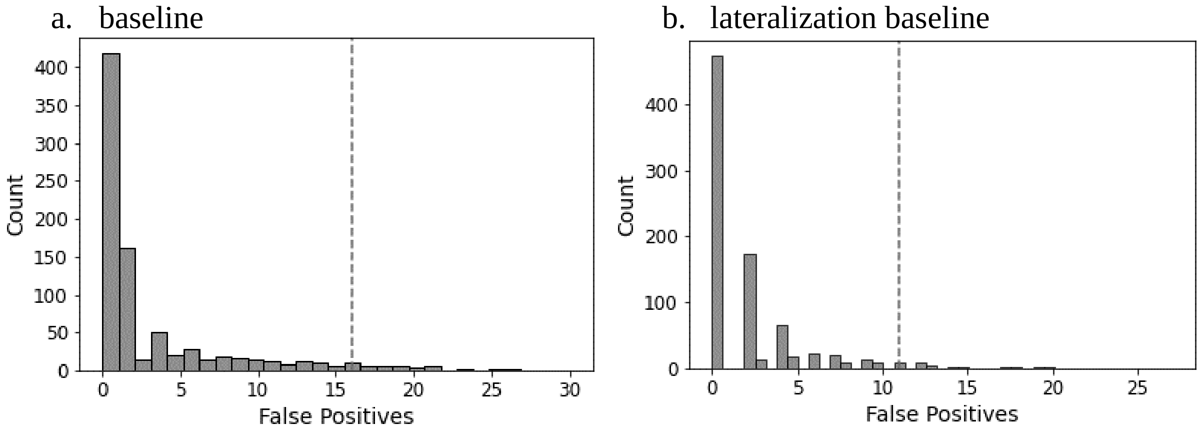
b. testing for lateralization

Note: Dataset size is measured in the total number of time-space units. For each data size the percentage of datasets that test significant was computed per condition. In both baseline conditions, the percentage of globally significant tests is above the desired value (FWER = 5%). The scatter plot has been fitted with a logarithmic regression function.

With the current critical p-value calculation, the STW approach leads to 2-30 false positive time-space units per dataset ($M_{baseline} = 7,\ SD_{baseline} = 6,\ M_{lat.\ base.} = 5,$ $SD_{lat.\ base.} = 4$). This is equivalent to a very low local type I ER ($M =\ 0.4\%,\ SD = 0.8\%$) for globally significant baseline conditions tested with the STW approach. Nevertheless, this only ensures that the STW approach provides GFWER control: the probability that "at least 15 type I errors" occur at the local level ($\alpha = 0.05$) (see Figure 8). Because of the successive time window criterion, the GFWER would ensure that there are no more than seven pairs (15 ÷ 2) of successive units.

Figure 8

Distribution of False Positive Count in Baseline Conditions Across Datasets



a. baseline

b. lateralization baseline

Note: The GFWER ensures that only 5% of datasets have at least the number of false positives exceeding the 95% quantile (dotted bar).

To ensure strict FWER correction, the critical p-value should be adjusted according to the specific dataset to be tested. The datasets where a higher number of local tests was performed (> 2000) have a lower critical p-value (see Figure 9). However, this lower critical p-value does not seem to be proportionate to the increase in dataset size, therefore global type I ER is not properly corrected for in large datasets. Besides the number of tested time-space units, the signal-to-noise ratio can affect local and global test results. As it cannot be established beforehand how noisy the data is, the critical p-value should be adjusted for each dataset separately, such that noise does not test significant. In practice, this can be done by first evaluating the baseline rest state condition (which should contain only noise), before testing for effects in conditions which are expected to contain signal. A wide range of critical p-values can be tested on the baseline, to establish which one is optimal for FWER control. To ensure that type II ERs will also stay low, the largest critical p-value (which still controls the FWER) should always be selected. Otherwise, too much sensitivity might be sacrificed.

Therefore, an iterative process can be used to determine the optimal critical *p*-value by slowly decreasing the value until no more local type I errors occur on the baseline condition (see Appendix E).

Discussion

This study aimed to validate the STW approach for solving the multiple testing problem in ERP research. The performance of the STW approach was quantified through specificity, sensitivity, and precision metrics, and compared against existing FWER control methods. The study also aimed to explore the impact of data dependencies, a priori assumptions, and the signal-to-noise ratio on the performance of the STW approach. Lastly, the study aimed to suggest improvements for the calculation of the critical *p*-value such that optimal performance is achieved across varying datasets.

Findings

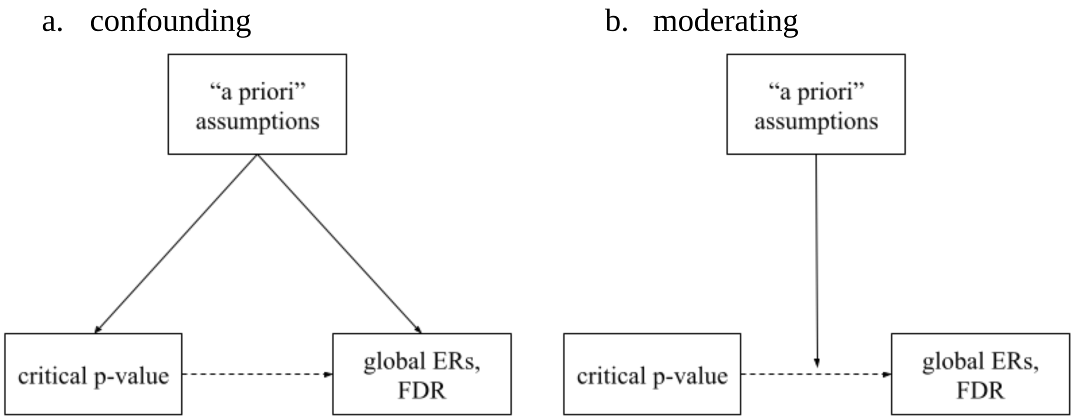
The validity of the STW approach was assessed according to the first five subgoals of this research. First, although very little type I errors occur at the local level (type I ER < 1%), there is a random chance that "at least one type I error" occurs when testing a baseline condition (FWER ~ 50%). The STW approach cannot identify the absence of signal or the absence of lateralization in baseline data (Subgoal I, IV). Therefore, this method does not offer strict FWER control. Nevertheless, it can provide better control for the GFWER: the probability that "at least 15 type I errors" occur at the local level. Next, although a lot of local effects are missed (type II ER ~ 60%), the STW approach can identify that a signal is present in the visual left/right stimulus presentation conditions, and that there is lateralization between the two conditions (Subgoals II, V). The identified local effects can be used to determine the time and location of the signal (with FDR ~ 30%) and the time and location of lateralization (with FDR < 10%). Therefore, the STW approach is a precise statistical method for identifying the presence of ERP effects, but it does not offer much sensitivity. Lastly, the STW approach can identify that there is a difference between the two visual stimuli presentation conditions (global type II ER). However, it cannot determine the precise difference (FDR = 43%) because most local effects are missed (type II ER = 98%). Therefore, it cannot be validated whether this method can identify the difference between two visual stimulus presentation conditions given the data used in this simulation (Subgoal IV).

For comparison (Subgoal VI), the CP method (Maris & Oostenveld, 2007) can control the FWER (<5%) in baseline data, it can identify the presence of signal and lateralization, and similar to the STW approach, it cannot identify the difference between visual stimulus conditions either. Nevertheless, the two methods are fundamentally very different. On the one hand, the CP method is less sensitive to noise at the local level, which leads to better FWER control, but it sacrifices precision. As data resolution gets lost in the cluster forming step, the precision within clusters cannot be determined. It cannot be said how many or which time-space units within a true positive cluster represent true local effects (Sassenhagen & Draschkow, 2019). On the other hand, the STW approach is susceptible to the presence of some local type I errors because of the high number of tests performed at the time-space unit level. However, this method provides higher overall precision. As long as the number of local type I errors stays low, the boundaries of an existing effect can be specified at the level of time-space units with the STW approach. Additionally, Bonferroni provides FWER similar to the CP method, and precision similar to the STW method, however it sacrifices too much sensitivity (local type II ER >87%). Therefore, the cluster-permutation method is better at identifying whether an effect is present or not, while the STW approach is better at identifying the boundaries of an effect. Bonferroni can only identify very high amplitude peaks of ERP components.

Furthermore, the performance of the STW was assessed for various data parameters (Subgoal VII). It is not clear whether the application of a mild bandpass filter (0.1, 30 *Hz*) (Tanner et al., 2015) impacted the performance of the STW approach, but all other data parameters clearly did. Firstly, a lower signal-to-noise ratio (determined by low amplitude values or high noise values) decreases the sensitivity of results and increases precision. This suggests that only high amplitude ERP peaks might be identified in data with low signal-to-noise ratio. Secondly, sensitivity was also lower when window size was larger. Groppe et al., (2011a) recommends down sampling EEG data, however this does not improve the performance of mass univariate tests if the effect does not span across a wider time range, like in the case of early ERP components simulated in this study. However, a window size that is too small can decrease both specificity and precision, as more false positives arise. A similar effect can be observed for the spatial dimension as well. Lower electrode density increases both sensitivity and precision of the STW approach. These results suggest that data with less dependencies (lower sampling rate or lower electrode density) is associated with higher precision. Lastly, the results of this study support existing findings that even broadly defined a priori assumptions increase the performance of mass univariate testing (Fields & Kuperberg, 2020). Time and location a priori highly increase sensitivity, precision, and specificity of the STW approach. A priori location assumptions had by far the greatest impact of all data parameters, but they cannot be used in the case of fully exploratory research (Groppe et al., 2011a).

Finally, the calculation of the critical *p*-value seems to allow too many false positives when the number of tested time-space units is high. FWER and FDR are positively correlated with dataset size. However, it is not clear if the critical *p*-value has a strong effect on performance, or if the observed correlations are affected by a third variable as well. A priori assumptions proved to influence the performance of the STW approach, by focusing testing on the time-space units where an effect is indeed expected. Therefore, a priori assumptions could influence the relation between the critical *p*-value calculation and performance metrics (global ERs, FDR) as a confounding or moderating variable (see Figure 9). A priori assumptions influence both how many local tests are performed, and exactly what time-space units are tested.

Figure 9
The Possible Effect of A Priori Assumptions on the Relation Between the Critical P-Value Calculation and Performance Metrics



Note: The use of a priori assumptions highly reduces dataset size. The smallest datasets, associated with higher performance are defined by the use of a priori assumptions. Therefore, the observed correlations between critical *p*-value (defined by dataset size) and performance metrics might be caused (or at least moderated) by the use of a priori assumptions.

Finally, to ensure that data size and the signal-to-noise ratio is accounted for, the *p*-value should be empirically adjusted based on rest state EEG before testing for ERP effects in experimental data. This introduces additional computation steps to the analysis of EEG data, but it ensures that the STW approach is tailored to the data at hand. An empirically selected *p*-value could improve STW performance and the interpretation of results. By adjusting the *p*-value to the noise level in baseline conditions, specificity and precision of results can be maximized. However, this does not solve the problem of reduced sensitivity.

Contributions to the Field of ERP Studies

Based on the results of this study, recommendations can be made for future exploratory ERP research. First, the STW approach has the potential of being a valid FWER control method for solving the multiple testing problem. If the critical *p*-value was empirically adjusted to the noise level in baseline data, the method could achieve very high specificity and precision. With the critical *p*-value calculation proposed by van der Lubbe et al. (2014, 2019) the method cannot ensure strict FWER control, but it yields better precision than the CP method (Maris & Oostenveld, 2007) and more sensitivity than the classic Bonferroni correction. A researcher should make a choice about which method to use depending on the aim of the ERP study, and the type of effect that is expected. The CP method is most suitable when the focus is on simply identifying the presence or absence of an effect at the global level. Bonferroni is most suitable for identifying high amplitude peaks of ERP components. The STW approach sacrifices specificity but can offer the sensitivity and precision needed for identifying the boundaries of ERP effects. Furthermore, if broad assumptions can be made about the time and especially location of the expected effect, they should be used to improve performance of the STW approach. The CP method could be used to first check whether an effect is present or not, and then use the STW approach to identify the specific boundaries of the effect on a priori selected electrodes from a significant cluster.

Furthermore, the methodology used in this study can provide a blueprint for extensive validation testing. The MNE Python library (Gramfort et al., 2013) proved to be a useful and complete tool for creating Monte Carlo simulations to measure the performance of statistical methods used in EEG analysis. Firstly, synthetic ERP signal can be freely designed from scratch, neural sources can be drawn from sample subject anatomy, and realistic forward modelling and noise can be used to generate scalp level EEG data. Secondly, the library provides extensive data processing options that allow the creation of a wide range of EEG datasets. Researchers can turn to similar methodology to thoroughly validate other existing methods under a wide range of data parameters. The methodology of this study can be fully reproduced (Paul et al., 2021), and future studies can be run by editing parameters in the source code (Coroiu, 2022).

Limitations

Firstly, this research was limited to comparing parametric methods against non-parametric methods only at the global level. Since the CP method (Maris & Oostenveld, 2007) cannot be used to make statements about the significance of individual data units, it is not possible to compare the sensitivity and precision of this method against the STW and Bonferroni approach at the local level. The precision calculated for the CP methods only measures the cluster-level precision (the number of significant clusters that contain at least one time-space unit with true signal). By definition, the precision within a cluster cannot be assessed. Sassenhagen & Draschkow (2019) assessed the temporal sensitivity of the CP method by taking the earliest time point within a cluster. This study did not extend to qualitatively assessing the identified boundaries of ERP effects. The metrics used in this research simply provide quantitative measures of error rates, but it is still unknown where exactly these errors occur and what specific consequences they have on the interpretation of results (e.g., by how much is an ERP effect underestimated?).

Secondly, the simulated signal was limited to the earlier components of the visual ERP. These components are associate with an attentional (orienting) effect (di Russo et al., 2003) and low-level processing of visual stimuli (Key et al., 2005; Woodman, 2010). These components are also characterized by shorter peaks, making the effect in the EEG data much more localized in time. In contrast, the later ERP components, associated with higher level processing (e.g., stimuli recognition) (Key et al., 2005; Woodman, 2010), have a much longer duration. Introducing later visual ERP components can change the performance of statistical methods. For example, the CP method is expected to be more suitable if the effect is broadly distributed over time (Groppe et al., 2011a).

Thirdly, it is not clear how to simulate fully synthetic and realistic EEG noise. The noise in this study was simulated based on the examples provided by MNE (Bekthi et al., 2022). However, other methods exist in the literature. Most commonly, real noise was obtained from experimental recordings of participants in resting state (Fields & Kuperberg, 2020; Groppe et al., 2011b). With this approach, a synthetic ERP is added on top of existing EEG data. However, when testing the FWER control, it is crucial to be certain that the recorded data does not contain any real ERP signal. To simulate baseline EEG data from scratch, other, more complex methods can be employed for creating realistic EEG noise (e.g., Barzegaran et al., 2019). As EEG noise is not fully understood, the noise simulated in this study cannot be clearly validated.

Lastly, this research was limited to the anatomy of one subject. The source location and tissue conductivity were based on the MNE sample head data. The locations of the ipsilateral and contralateral source were very close together. Therefore, the simulated EEG data in all the 20 simulated subjects presented reduced lateralization. This highly influenced the results when testing the difference between right/left visual stimulus presentation conditions. The activation patterns from the two conditions were very similar, making it difficult for all methods to distinguish between them. Furthermore, Woodman (2010) highlights that individual differences in ERP arise from differences in brain tissue. In this study, some variation was introduced in the source signal for each simulated participant by modifying the generated ERP wave, however more realistic variation between participants can only be introduced by using different head models. Additionally, differences in IHTT delay between participants were observed in previous research (Moes et al., 2007), but not modelled in this simulation.

Recommendations for Future Research

The multiple testing problem can be considered a binary classification problem. Statistical methods for EEG data analysis used in ERP research aim to classify each data unit (or cluster) under one of two labels: signal or noise. Therefore, validation studies in this field can benefit from the methodology used in validation studies in the field of machine learning. Existing simulation studies used for assessing the performance of FWER methods (Fields & Kuperberg, 2020; Groppe et al., 2011b) are focused on concepts from classical hypothesis testing: type I and type II error rates. However, other metrics can be used to quantitatively assess how well a method performs in discriminating signal from noise. The traditional F_1 score can be a much more useful metric for assessing the balance between sensitivity and precision. Accuracy should not be used, as it is not guaranteed that the number of data units containing signal is equal to the number of data units containing noise. Additionally, the Matthews correlation coefficient (Chicco & Jurman, 2020) has been recently proposed as a better alternative to the F_1 score for imbalanced data.

Qualitative assessment can be performed for comparing the performance of the STW approach to other existing methods. Future research can focus on evaluating how well statistical methods estimate the onset and offset of an ERP, and how well they can localize the effect in space. Sassenhagen & Draschkow (2019) highlight that different implementations of the cluster-permutation method offer varying degrees of precision for the spatial and temporal dimension of results. Additionally, future research could evaluate how precisely statistical methods can identify the separate ERP components. An important aspect to investigate is how spatial and temporal smearing affects the identification of ERP components, as the signal coming from different ERP peaks can blend at the scalp level (Burle et al., 2015).

The proposed critical *p*-value baseline adjustment can be tested using the data and metrics from this study. The calculated critical *p*-value will depend on the noise level present in baseline conditions. With high noise datasets it might sacrifice too much sensitivity, behaving similarly to Bonferroni. Future research can evaluate how the noise level and critical *p*-value affect the performance of the STW. For visualization, Receiver Operating Characteristic (ROC) curves could be used to plot the binary classification ability of the STW window as the critical *p*-value varies. Additionally, interaction effects between the different data parameters can be assessed in future research. This can help determine how big the effect of data size and dependencies (defined by time window size and electrode density) is compared to the effect of using a priori assumptions.

The STW approach relies on the assumption of data independence, which is not correct. Underlying properties of the data are not accounted for in the parametric tests performed for this method. Data dependencies introduced by spatial and temporal smearing at the scalp level (Burle et al., 2015) can decrease the precision of results. Furthermore, data dependencies exist at the participant and trial level (latency and amplitude for the simulated EEG data in this study). However, trial level variation gets lost with averaging of data points into time-space units per participant, and participant level variation gets lost with the parametric test performed. Bayesian statistical methods that factor in all these data dependencies might prove more suitable for identifying the presence and boundaries of ERP effects in EEG data (Wu et al., 2016). The performance of Bayesian methods can be evaluated in future research following the same simulation methodology.

More extensive Monte Carlos simulations can be created using a wider range of data parameters and data parameter values. For example, the same research could be performed for peripheral auditory stimulus conditions, such that the source dipole signal would be located more laterally within the brain. ERPs of various durations (containing components of different shapes) can be simulated, to evaluate the performance of methods in recognizing broader effects. Realistic EEG data can be simulated using real baseline noise, or other computational techniques for generating synthetic noise. Furthermore, with different head models, the performance of the STW approach and other methods can be assessed for testing differences between participant groups.

Conclusion

This study evaluated the performance of the STW approach against the CP method and the Bonferroni correction for identifying ERP effects. Both the CP method and Bonferroni offer better FWER control, but the STW approach can offer more sensitivity and precision for identifying ERP effects. Even if they are broadly defined, a priori assumptions about the location of the expected effects should be used as they can improve the performance of the STW approach. In a fully exploratory setting, the CP method could be used to define a priori assumptions about an effect before using the STW method. For better FWER control, the critical *p*-value could be adjusted according to baseline data. To evaluate the performance of the STW in identifying the boundaries of ERP effects, qualitative assessment of the results should be performed. Furthermore, the field of ERP research could benefit from using newer techniques like Bayesian methods that can account for data dependencies or using metrics from machine learning to assess the performance of statistical methods. Lastly, future research can use the methodology of this study to create simulated EEG data and evaluate the performance of mass univariate statistical methods under more data parameters.

Glossary

This glossary summarizes the definitions of statistical metrics used to evaluate the performance of statistical methods employed in this study (see Table 5).

Table 5

Glossary of Statistical Metrics

Term	Definition
False Discovery Rate (FDR)	the rate of false positives in all positive (significant) results
false negative or type II error	a test which was expected to be positive (effect present), but resulted negative (no significant effect)
false negative rate or type II error rate	the rate of false negatives in all tests expected positive (effects present); the probability of missing an effect
false positive or type I error	a test which was expected to be negative (effect not present), but resulted positive (significant effect)
false positive rate or type I error rare	the rate of false positives in all tests expected negative (effects not present); the probability of identifying a "bogus effect"
Family Wise Error Rate (FWER)	the probability of having at least one false positive result in a "family" of simultaneous tests; the rate of tested datasets which contain at least one false positive data units/clusters
Generalized Family Wise Error Rate (GFWER)	the probability of having at least p false positive results in a "family" of simultaneous tests; the rate of tested datasets which contain at least p false positive data units/clusters (*with $p \epsilon N$, $p < n$, where n is the total number of data units/clusters in the dataset*)
global type I error	a dataset which was expected to be negative (conditions with no effects present: baseline data) but resulted positive (there was at least one significant data unit/cluster in the dataset)
global type I error rate	the rate of false positive datasets in all datasets expected negative (conditions with no effects present: baseline data); equivalent to FWER
global type II error	a dataset which was expected to be positive (conditions where effects are present: right/left visual stimulus, difference, lateralization), but resulted negative (there was no significant data unit/cluster in the dataset)
global type II error rate	the rate of false negative datasets in all datasets expected positive (conditions where effects are present: right/left visual stimulus, difference, lateralization)
local type I error	a data unit which was expected to be negative (time and location do not correspond to a true effect), but resulted in a positive (a significant effect at this combination of time and space)
local type I error rate	the rate of false positive data units in all data units expected negative (time and location do not correspond to a true effect)

local type II error	a data unit which was expected to be positive (time and location correspond to a true effect), but resulted in a positive (no significant effect at this combination of time and space)
local type II error rate	the rate of false positive data units in all data units expected positive (time and location correspond to a true effect)
precision	the rate of true positives in all positive results; the complement of FDR (the lower the FDR, the higher the precision)
resolution	the number of units used to represent the dataset; EEG temporal resolution is determined by the sampling rate (time window size), while EEG spatial resolution is determined by electrodes density (the number of electrodes used); the cluster-based permutation method reduces the initial resolution of EEG data to the number of formed clusters
sensitivity	the rate of true positives in all tests expected positive; the complement of type II error rate (the lower the type II ER, the higher the sensitivity)
specificity	the rate of true negatives in all tests expected negative; the complement of type I error rate (the lower the type I ER, the higher the specificity)

References

Baillet, S., Mosher, J. C., & Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, *18*(6), 14–30. https://doi.org/10.1109/79.962275

Barzegaran, E., Bosse, S., Kohler, P. J., & Norcia, A. M. (2019). EEGSourceSim: A framework for realistic simulation of EEG scalp data using MRI-based forward models and biologically plausible signals and noise. *Journal of Neuroscience Methods*, *328*. https://doi.org/10.1016/j.jneumeth.2019.108377

Bekthi, Y., Wronkiewicz, M., & Larson, E. (2022). *Generate Simulated Raw Data*. MNE Data Simulation. https://mne.tools/stable/auto_examples/simulation/simulate_raw_data.html#sphx-glr-auto-examples-simulation-simulate-raw-data-py

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210–220. https://doi.org/10.1016/j.ijpsycho.2015.05.004

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Clarke, S., & Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Annals of Statistics*, *37*(1), 332–358. https://doi.org/10.1214/07-AOS557

Coroiu, A. (2022). *Testing the Successive Time Window Approach with Simulated EEG Data*. https://github.com/AlexCoroiu/window_correction

di Russo, F., Martínez, A., & Hillyard, S. A. (2003). Source Analysis of Event-related Cortical Activity during Visuo-spatial Attention. *Cerebral Cortex*, *13*(5), 486–499. https://doi.org/10.1093/cercor/13.5.486

di Russo, F., Martínez, A., Sereno, M. I., Pitzalis, S., & Hillyard, S. A. (2002). Cortical sources of the early components of the visual evoked potential. *Human Brain Mapping*, *15*(2), 95–111. https://doi.org/10.1002/hbm.10010

Fields, E. C., & Kuperberg, G. (2020). Having your cake and eating it too: Flexibility and power with mass univariate statistics for ERP data Title: Mass univariate stats for ERP data. *Psychophysiology*, *57*(2). https://doi.org/10.1111/psyp.13468

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

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

Hallez, H., Vanrumste, B., Grech, R., Muscat, J., de Clercq, W., Vergult, A., D'Asseler, Y., Camilleri, K. P., Fabri, S. G., van Huffel, S., & Lemahieu, I. (2007). Review on solving the forward problem in EEG source analysis. In *Journal of NeuroEngineering and Rehabilitation* (Vol. 4). https://doi.org/10.1186/1743-0003-4-46

Key, A. P. F., Dove, G. O., & Maguire, M. J. (2005). Linking Brainwaves to the Brain: an ERP primer. *Developmental Neuropsychology*, *27*(2), 183–215. https://doi.org/10.1207/s15326942dn2702_1

Kim, K. I., & van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics, 9*. https://doi.org/10.1186/1471-2105-9-114

Klem, G. H., Otto Lu Èders, H., Jasper, H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and Clinical Neurophysiology. Supplement*, *52*, 3–6.

Liu, A. K., Dale, A. M., & Belliveau, J. W. (2002). Monte Carlo simulation studies of EEG and MEG localization accuracy. *Human Brain Mapping*, *16*(1), 47–62. https://doi.org/10.1002/hbm.10024

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. 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), 146–157. https://doi.org/10.1111/psyp.12639

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

Moes, P. E., Brown, W. S., & Minnema, M. T. (2007). Individual differences in interhemispheric transfer time (IHTT) as measured by event related potentials. *Neuropsychologia*, *45*(11), 2626–2630.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Paul, M., Govaart, G. H., & Schettino, A. (2021). Making ERP research more transparent: Guidelines for preregistration. *International Journal of Psychophysiology, 164*, 52–63. https://doi.org/10.1016/j.ijpsycho.2021.02.016

Rauss, K., Schwartz, S., & Pourtois, G. (2011). Top-down effects on early visual processing in humans: A predictive coding framework. In *Neuroscience and Biobehavioral Reviews* (Vol. 35, Issue 5, pp. 1237–1253). https://doi.org/10.1016/j.neubiorev.2010.12.011

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6). https://doi.org/10.1111/psyp.13335

Takemura, H., Yuasa, K., & Amano, K. (2020). Predicting neural response latency of the human early visual cortex from MRI-based tissue measurements of the optic radiation. *ENeuro, 7*(4), 1–18. https://doi.org/10.1523/ENEURO.0545-19.2020

Talsma, D., Wijers, A. A., Klaver, P., & Mulder, G. (2001). Working memory processes show different degrees of lateralization: Evidence from event-related potentials. *Psychophysiology*, *38*(3), 425–439. https://doi.org/10.1111/1469-8986.3830425

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Tanner, D., Norton, J. J. S., Morgan-Short, K., & Luck, S. J. (2016). On high-pass filter artifacts (they're real) and baseline correction (it's a good idea) in ERP/ERMF analysis. In *Journal of Neuroscience Methods* (Vol. 266, pp. 166–170). Elsevier B.V. https://doi.org/10.1016/j.jneumeth.2016.01.002

Teplan, M. (2002). Fundamentals of EEG Measurement. *Measurement Science Review*, *2*(2).

van der Lubbe, R. H. J., Bundt, C., & Abrahamse, E. L. (2014). Internal and external spatial attention examined with lateralized EEG power spectra. *Brain Research, 1583*(1), 179–192. https://doi.org/10.1016/j.brainres.2014.08.007

van der Lubbe, R. H. J., de Kleine, E., & Rataj, K. (2019). Dyslexic individuals orient but do not sustain visual attention: Electrophysiological support from the lower and upper alpha bands. *Neuropsychologia*, *125*, 30–41. https://doi.org/10.1016/j.neuropsychologia.2019.01.013

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception & Psychology*, *72*(8), 2031–2046. https://doi.org/10.3758/APP.72.8.2031

Wu, W., Nagarajan, S., & Chen, Z. (2016). Bayesian Machine Learning: EEG\/MEG signal processing measurements. *IEEE Signal Processing Magazine*, *33*(1), 14–36. https://doi.org/10.1109/MSP.2015.2481559

Appendix A

Example Processing, Analysis and Results Data

Table A1

Example of Prepared Data

part	time (ms)	electrode	voltage (µV)
1	0	C3	0.185332
1	0	C4	0.570573
1	0	CP3	0.456645
1	0	CP4	0.074455
1	0	CPz	-0.42881
1	0	Cz	-0.16973
1	0	F3	1.353519
1	0	F4	-0.27631
1	0	F7	0.586506
1	0	F8	0.415466
1	0	FC3	0.713648
1	0	FC4	0.690223
1	0	FCz	0.615512
1	0	FT7	0.532583
…	…	…	…
20	500	F8	-0.54578
20	500	FC3	-0.54716
20	500	FC4	0.282355
20	500	FCz	0.871459
20	500	FT7	0.470344
20	500	FT8	0.517057
20	500	Fp1	-0.63602
20	500	Fp2	-0.34879
20	500	Fpz	0.814388
20	500	Fz	-0.19965
20	500	O1	0.844423
20	500	O2	0.452424
20	500	Oz	0.337198
20	500	P3	-0.85392
20	500	P4	0.605701
20	500	P7	-0.62579
20	500	P8	-0.41356
20	500	Pz	0.990291
20	500	T7	0.536308
20	500	T8	-0.91903
20	500	TP7	-0.11419
20	500	TP8	-1.83971

Note: For each participant, an amplitude value was stored for every time-space unit. These participant level values were obtained by averaging data across trials. The data here was prepared for testing the visual left condition

Table A2

Example Analysed Data with the Successive Time Window Approach

time (ms)	electrode	*p*-val	crit. *p*-val	crit *p*-val reject	window reject
…	…	…	…	…	…
144	Fpz	0.011706	0.004365	FALSE	FALSE
156	Fpz	0.00615	0.004365	FALSE	FALSE
168	Fpz	0.004139	0.004365	TRUE	FALSE
180	Fpz	0.049426	0.004365	FALSE	FALSE
192	Fpz	0.713878	0.004365	FALSE	FALSE
…	…	…	…	…	…
108	POz	0.961938	0.004365	FALSE	FALSE
120	POz	0.142793	0.004365	FALSE	FALSE
132	POz	0.003601	0.004365	TRUE	TRUE
144	POz	0.001023	0.004365	TRUE	TRUE
156	POz	0.000026	0.004365	TRUE	TRUE
168	POz	0.000109	0.004365	TRUE	TRUE
180	POz	0.03771	0.004365	FALSE	FALSE
…	…	…	…	…	…

Note: A two-tailed *t*-test was performed at each time-space unit using the calculated critical *p*-value for the prepared dataset (here window size = 12 ms, electrode density = 64, no a priori assumptions, testing condition = visual left). The results of the *t*-test were then corrected based on the successive time window criterion. The results in the final column represent the significance for each time-space unit.

Table A3

Example Analysed Data with the Classic Bonferroni Correction

time (ms)	electrode	*p*-val	crit. *p*-val	crit *p*-val reject
…	…	…	…	…
144	Fpz	0.011706	0.0000186	FALSE
156	Fpz	0.00615	0.0000186	FALSE
168	Fpz	0.004139	0.0000186	FALSE
180	Fpz	0.049426	0.0000186	FALSE
192	Fpz	0.713878	0.0000186	FALSE
…	…	…	…	…
108	POz	0.961938	0.0000186	FALSE
120	POz	0.142793	0.0000186	FALSE
132	POz	0.003601	0.0000186	FALSE
144	POz	0.001023	0.0000186	FALSE
156	POz	0.000026	0.0000186	FALSE
168	POz	0.000109	0.0000186	FALSE
180	POz	0.03771	0.0000186	FALSE
…	…	…	…	…

Note: A two-tailed *t*-test was performed at each time-space unit using the Bonferroni critical *p*-value for the prepared dataset (here window size = 12 ms, electrode density = 64, no a priori assumptions, testing condition = visual left). The results in the final column represent the significance for each time-space unit.

Table A4

Example Analysed Data with the Cluster Permutation Method

cluster	data units	*p*-val	crit. *p*-val	significant
1	[[12, 'T9']]	1	0.05	FALSE
2	[[36, 'AF8']]	1	0.05	FALSE
3	[[36, 'PO4'], [48, 'PO4']]	1	0.05	FALSE
4	[[36, 'Pz']]	1	0.05	FALSE
…	…	…	…	…
14	[[228,'CP1'], [192,'CP2'], [204,'CP2'], [216, 'CP2'], [240, 'CP3'], [204, 'CP4'], [204, 'CP5'], [216, 'CP5'], [228, 'CPz'], [240, 'CPz'], [252, 'CPz'], [204, 'Iz'], [216, 'Iz'], [228, 'Iz'], [240, 'Iz'], [204, 'O1'], [216, 'O1'], [228, 'O1'], [240, 'O1'], [204, 'O2'], [216, 'O2'], [228, 'O2'], [240, 'O2'], [252, 'O2'], [204, 'Oz'], [216, 'Oz'], [228, 'Oz'], [240, 'Oz'], [204, 'P1'], [216, 'P1'], [228, 'P1'], [240, 'P1'], [204, 'P2'], [216, 'P2'], [228, 'P2'], [240, 'P2'], [204, 'P3'], [216, 'P3'], [228, 'P3'], [240, 'P3'], [252, 'P3'], [204, 'P4'], [216, 'P4'], [228, 'P4'], [240, 'P4'], [252, 'P4'], [204, 'P5'], [216, 'P5'], [228, 'P5'], [240, 'P5'], [252, 'P5'], [204, 'P6'], [216, 'P6'], [228, 'P6'], [204, 'P7'], [216, 'P7'], [240, 'P7'], [192, 'P8'], [204, 'P8'], [204, 'PO3'], [216, 'PO3'], [228, 'PO3'], [240, 'PO3'], [204, 'PO4'], [216, 'PO4'], [228, 'PO4'], [240, 'PO4'], [216, 'PO7'], [228, 'PO7'], [204, 'PO8'], [216, 'PO8'], [228, 'PO8'], [204, 'POz'], [216, 'POz'], [228, 'POz'], [240, 'POz'], [204, 'Pz'], [216, 'Pz'], [228, 'Pz'], [240, 'Pz']]	0.001953	0.05	TRUE
…	…	…	…	…
41	[[36, 'AF4']]	1	0.05	FALSE
42	[[48, 'C4'], [48, 'C6'], [60, 'C6']]	0.996094	0.05	FALSE
43	[[48, 'CPz'], [60, 'CPz'], [72, 'CPz']]	0.984375	0.05	FALSE
44	[[48, 'F7']]	1	0.05	FALSE
…	…	…	…	…

Note: Data units with significant *t*-values were clustered according to temporal and spatial proximity. A *p*-value was calculated for each created. Clusters with *p*-values smaller than the critical *p*-value are considered significant. The results presented here come from a dataset prepared with window size = 12 ms, electrode density = 64, no a priori assumptions, testing condition = visual left.

Table A5

Example Results Successive Time Window Approach

window size	time a priori	electrode density	location a priori	condition	crit *p*-val	total	positives	global significant	TP	FP	TN	FN	Type I ER	Type II ER
4	TRUE	86	TRUE	baseline	0.005324	1764	0	FALSE	0	0	1764	0	0	0
4	TRUE	86	TRUE	visual right	0.005324	1764	345	TRUE	337	8	692	727	0.011429	0.683271
4	TRUE	86	TRUE	visual left	0.005324	1764	346	TRUE	346	0	700	718	0	0.674812
4	TRUE	86	TRUE	difference	0.005324	1764	7	FALSE	5	2	698	1059	0.002857	0.995301
4	TRUE	86	FALSE	baseline	0.003038	5418	14	FALSE	0	14	5404	0	0.002584	0
4	TRUE	86	FALSE	visual right	0.003038	5418	419	TRUE	300	119	4235	764	0.027331	0.718045
4	TRUE	86	FALSE	visual left	0.003038	5418	430	TRUE	289	141	4213	775	0.032384	0.728383
4	TRUE	86	FALSE	difference	0.003038	5418	12	FALSE	3	9	4345	1061	0.002067	0.99718
4	TRUE	64	TRUE	baseline	0.00664	1134	0	FALSE	0	0	1134	0	0	0
4	TRUE	64	TRUE	visual right	0.00664	1134	274	TRUE	263	11	439	421	0.024444	0.615497
4	TRUE	64	TRUE	visual left	0.00664	1134	254	TRUE	254	0	450	430	0	0.628655
4	TRUE	64	TRUE	difference	0.00664	1134	10	FALSE	6	4	446	678	0.008889	0.991228
4	TRUE	64	FALSE	baseline	0.003521	4032	13	FALSE	0	13	4019	0	0.003224	0
4	TRUE	64	FALSE	visual right	0.003521	4032	320	TRUE	232	88	3260	452	0.026284	0.660819
4	TRUE	64	FALSE	visual left	0.003521	4032	341	TRUE	211	130	3218	473	0.038829	0.69152
4	TRUE	64	FALSE	difference	0.003521	4032	7	FALSE	3	4	3344	681	0.001195	0.995614
4	TRUE	31	TRUE	baseline	0.00996	504	0	FALSE	0	0	504	0	0	0
4	TRUE	31	TRUE	visual right	0.00996	504	128	TRUE	116	12	188	188	0.06	0.618421
4	TRUE	31	TRUE	visual left	0.00996	504	119	TRUE	119	0	200	185	0	0.608553
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

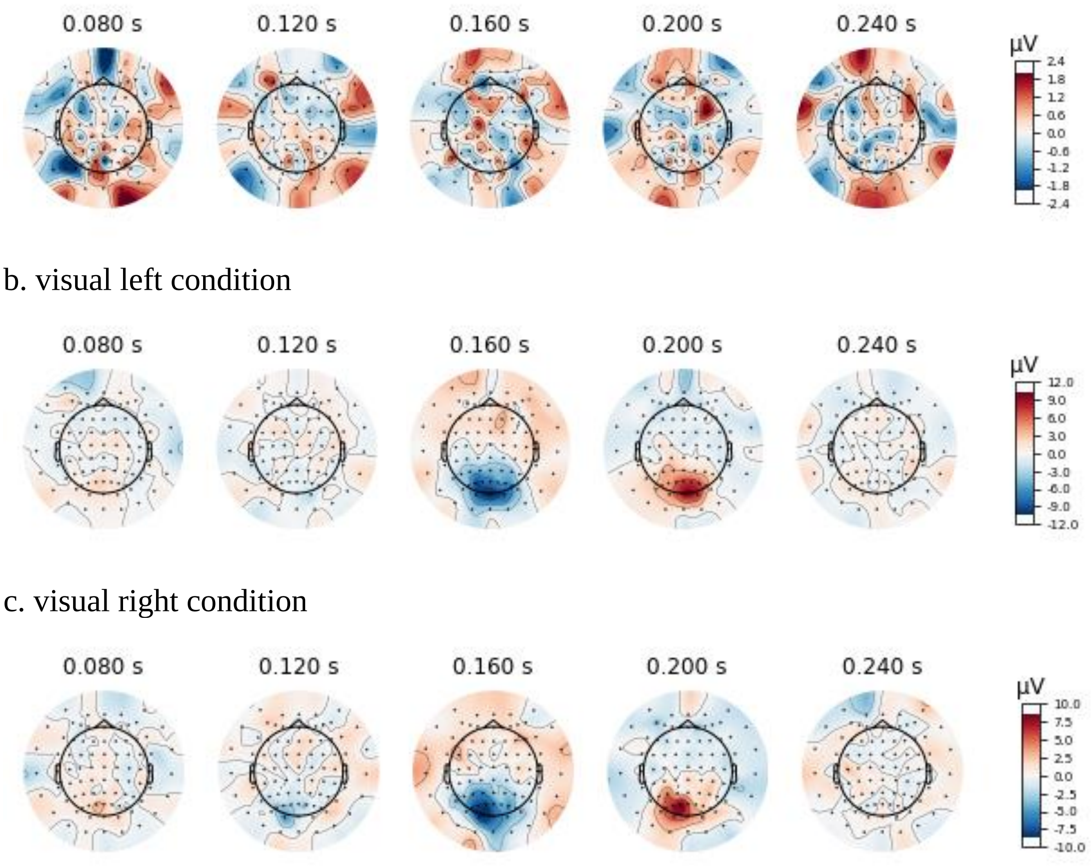
Note: The confusion matrix values and subsequent performance metrics (type I, II ER, FDR) are calculated for each dataset according to the results of the different analysis method. A dataset is defined by the data parameters used to simulate and process the synthetic EEG recordings. Confusion Matrices are calculated per condition.

Appendix B

ERP visualisation

Figure B1

Topographic Map for Participant 1



a. baseline condition

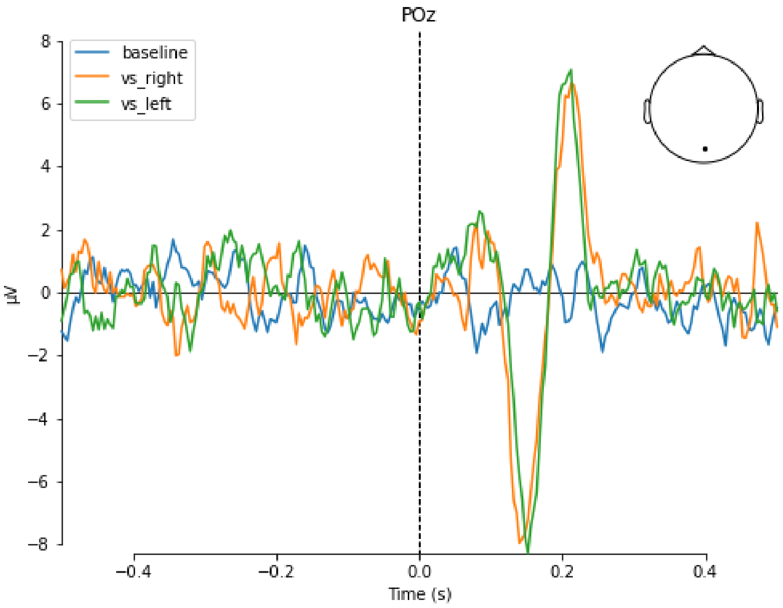
b. visual left condition

c. visual right condition

Note: Values for each electrode are averaged over trials. This data was generated under the following values for data parameters: (60,20) amplitude, high noise, and no band pass filter applied. Negativity is plotted downwards.

Figure B2

Average ERP at the POz electrode for Participant 1



Note: Values for each time point are averaged over trials. This data was generated under the following values for data parameters: (60,20) amplitude, high noise, no band pass filter applied. Negativity is plotted downwards.

Appendix C

The Impact of Data Parameters on Performance Metrics

Observed differences between parameter values (v) have been highlighted for local level type I/II ERs ($v_i - v_j > 5\%$), global level type I/II ERs ($v_j - v_j > 10\%$), and for FDR ($v_i - v_j > 10\%$).

Amplitude

Table C1

Mean Local ERs by Amplitude Value

condition	metric	(40, 20)	(60, 30)	(60, 20)	(80, 40)	(80, 30)	(80, 20)
baseline	type II ER	-	-	-	-	-	-
	type I ER	0.0028	0.0025	0.0009	0.0011	0.0029	0.002
visual left	type II ER	0.5968	0.5743	0.5576	0.5536	0.5397	0.5328
	type I ER	0.0273	0.0379	0.0463	0.0471	0.0554	0.0622
visual right	type II ER	0.6028	0.5856	0.5632	0.5611	0.5536	0.5377
	type I ER	0.0447	0.0615	0.072	0.08	0.0837	0.0936
difference	type II ER	0.9959	0.9855	0.9946	0.9714	0.9819	0.9888
	type I ER	0.0023	0.0035	0.0026	0.0041	0.005	0.003
lateralization baseline	type II ER	-	-	-	-	-	-
	type I ER	0.0025	0.0034	0.0024	0.0038	0.0047	0.0016
lateralization visual stimuli	type II ER	0.8791	0.7837	0.8419	0.7885	0.7521	0.719
	type I ER	0.0051	0.0062	0.0093	0.0135	0.0075	0.0089

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C2

Global ERs by Amplitude Value

condition	metric	(40, 20)	(60, 30)	(60, 20)	(80, 40)	(80, 30)	(80, 20)
baseline	type II ER	-	-	-	-	-	-
	type I ER	0.5972	0.3681	0.5764	0.5208	0.6111	0.4167
visual left	type II ER	0	0	0	0	0	0
	type I ER	-	-	-	-	-	-
visual right	type II ER	0	0	0	0	0	0
	type I ER	-	-	-	-	-	-
difference	type II ER	0.3681	0.3542	0.1528	0.3125	0.0833	0.0278
	type I ER	-	-	-	-	-	-
lateralization baseline	type II ER	-	-	-	-	-	-
	type I ER	0.4444	0.4167	0.4444	0.5208	0.4931	0.3958
lateralization visual stimuli	type II ER	0.0694	0	0	0	0	0
	type I ER	-	-	-	-	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C3

Mean FDR by Amplitude Value

condition	(40, 20)	(60, 30)	(60, 20)	(80, 40)	(80, 30)	(80, 20)
baseline	-	-	-	-	-	-
visual left	0.2115	0.2801	0.2597	0.3169	0.3061	0.2829
visual right	0.2469	0.3275	0.3176	0.3701	0.3558	0.3531
difference	0.6667	0.5429	0.3814	0.3514	0.4528	0.3114
lat. baseline	-	-	-	-	-	-
lat. visual stimuli	0.0975	0.0993	0.0808	0.0956	0.0926	0.0941

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Noise

Table C4

Mean Local ERs by Noise Value

condition	metric	high	low
baseline	type II ER	-	-
	type I ER	0.0024	0.0017
visual left	type II ER	0.5911	0.5272
	type I ER	0.0298	0.0623
visual right	type II ER	0.606	0.5287
	type I ER	0.0436	0.1016
difference	type II ER	0.9888	0.984
	type I ER	0.0024	0.0045
lateralization baseline	type II ER	-	-
	type I ER	0.0025	0.0037
lateralization visual stimuli	type II ER	0.8247	0.7634
	type I ER	0.0047	0.0121

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C5

Global ERs by Noise Value

condition	metric	high	low
baseline	type II ER	-	-
	type I ER	0.5694	0.4606
visual left	type II ER	0	0
	type I ER	-	-
visual right	type II ER	0	0
	type I ER	-	-
difference	type II ER	0.2546	0.1782
	type I ER	-	-
lateralization baseline	type II ER	-	-
	type I ER	0.4005	0.5046
lateralization visual stimuli	type II ER	0.0231	0
	type I ER	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C6

Mean FDR by Noise Value

condition	high	low
baseline	-	-
visual left	0.232	0.3204
visual right	0.2711	0.3859
difference	0.4275	0.4456
lat. baseline	-	-
lat. visual stimuli	0.0787	0.1075

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Band Pass

Table C7

Mean Local ERs by Band Pass Value

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.0011	0.003
visual left	type II ER	0.568	0.5503
	type I ER	0.0443	0.0477
visual right	type II ER	0.5682	0.5665
	type I ER	0.0703	0.0749
difference	type II ER	0.9863	0.9864
	type I ER	0.0032	0.0037
lateralization baseline	type II ER	-	-
	type I ER	0.0033	0.0029
lateralization visual stimuli	type II ER	0.7997	0.7884
	type I ER	0.007	0.0098

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C8

Global ERs by Band Pass Value

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.4491	0.581
visual left	type II ER	0	0
	type I ER	-	-
visual right	type II ER	0	0
	type I ER	-	-
difference	type II ER	0.2315	0.2014
	type I ER	-	-
lateralization baseline	type II ER	-	-
	type I ER	0.4606	0.4444
lateralization visual stimuli	type II ER	0.0231	0
	type I ER	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C9

Mean FDR by Band Pass Value

condition	FALSE	TRUE
baseline	-	-
visual left	0.2751	0.2773
visual right	0.321	0.336
difference	0.41	0.4629
lat. baseline	-	-
lat. visual stimuli	0.0779	0.1084

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Window Size

Table C10

Mean Local ERs by Window Size Value (s)

condition	metric	0.004	0.012	0.02
baseline	type II ER	-	-	-
	type I ER	0.0035	0.0016	0.0011
visual left	type II ER	0.5691	0.5061	0.6022
	type I ER	0.0542	0.0503	0.0336
visual right	type II ER	0.5716	0.5296	0.6008
	type I ER	0.0727	0.0803	0.0647
difference	type II ER	0.9849	0.983	0.9912
	type I ER	0.0052	0.0037	0.0014
lateralization baseline	type II ER	-	-	-
	type I ER	0.0042	0.0027	0.0023
lateralization visual stimuli	type II ER	0.8063	0.7765	0.7994
	type I ER	0.0119	0.0068	0.0066

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C11

Global ERs by Window Size Value (s)

condition	metric	0.004	0.012	0.02
baseline	type II ER	-	-	-
	type I ER	0.8819	0.4618	0.2014
visual left	type II ER	0	0	0
	type I ER	-	-	-
visual right	type II ER	0	0	0
	type I ER	-	-	-
difference	type II ER	0.0104	0.1424	0.4965
	type I ER	-	-	-
lateralization baseline	type II ER	-	-	-
	type I ER	0.7882	0.3785	0.191
lateralization visual stimuli	type II ER	0	0	0.0347
	type I ER	-	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C12

Mean FDR by Window Size Value (s)

condition	0.004	0.012	0.02
baseline	-	-	-
visual left	0.3007	0.2824	0.2455
visual right	0.3356	0.341	0.3089
difference	0.5075	0.4282	0.3133
lat. baseline	-	-	-
lat. visual stimuli	0.1551	0.0701	0.0533

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Time Interval

Table C13

Mean Local ERs by Time Interval Value (a priori)

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.0017	0.0024
visual left	type II ER	0.5769	0.5413
	type I ER	0.0262	0.0658
visual right	type II ER	0.579	0.5557
	type I ER	0.037	0.1082
difference	type II ER	0.9893	0.9834
	type I ER	0.0022	0.0047
lateralization baseline	type II ER	-	-
	type I ER	0.0025	0.0036
lateralization visual stimuli	type II ER	0.8064	0.7817
	type I ER	0.0043	0.0126

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C14

Global ERs by Time Interval Value (a priori)

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.5694	0.4606
visual left	type II ER	0	0
	type I ER	-	-
visual right	type II ER	0	0
	type I ER	-	-
difference	type II ER	0.2477	0.1852
	type I ER	-	-
lateralization baseline	type II ER	-	-
	type I ER	0.4907	0.4144
lateralization visual stimuli	type II ER	0.0116	0.0116
	type I ER	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C15

Mean FDR by Time Interval Value (a priori)

condition	FALSE	TRUE
baseline	-	-
visual left	0.2742	0.2782
visual right	0.3214	0.3356
difference	0.5268	0.3541
lat. baseline	-	-
lat. visual stimuli	0.1	0.0866

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Electrodes Density

Table C16

Mean Local ERs by Electrodes Density Value (nr. electrodes)

condition	metric	31	64	86
baseline	type II ER	-	-	-
	type I ER	0.003	0.0018	0.0014
visual left	type II ER	0.5337	0.5444	0.5993
	type I ER	0.0534	0.0443	0.0404
visual right	type II ER	0.5488	0.5488	0.6044
	type I ER	0.0866	0.069	0.0622
difference	type II ER	0.9805	0.9895	0.989
	type I ER	0.0044	0.0034	0.0025
lateralization baseline	type II ER	-	-	-
	type I ER	0.0044	0.0027	0.0021
lateralization visual stimuli	type II ER	0.7619	0.7916	0.8287
	type I ER	0.0146	0.0067	0.004

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C17

Global ERs by Electrodes Density Value (nr. electrodes)

condition	metric	31	64	86
baseline	type II ER	-	-	-
	type I ER	0.4688	0.5382	0.5382
visual left	type II ER	0	0	0
	type I ER	-	-	-
visual right	type II ER	0	0	0
	type I ER	-	-	-
difference	type II ER	0.2465	0.2118	0.191
	type I ER	-	-	-
lateralization baseline	type II ER	-	-	-
	type I ER	0.3819	0.4722	0.5035
lateralization visual stimuli	type II ER	0.0069	0.0139	0.0139
	type I ER	-	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C18

Mean FDR by Electrodes Density Value (nr. electrodes)

condition	31	64	86
baseline	-	-	-
visual left	0.3037	0.2714	0.2535
visual right	0.3624	0.3186	0.3045
difference	0.399	0.4783	0.4321
lat. baseline	-	-	-
lat. visual stimuli	0.1301	0.0866	0.063

Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Electrodes Location

Table C19

Mean Local ERs by Electrodes Location Value (a priori)

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.0014	0.0028
visual left	type II ER	0.5909	0.5274
	type I ER	0.0835	0.0086
visual right	type II ER	0.5869	0.5478
	type I ER	0.0957	0.0495
difference	type II ER	0.9912	0.9815
	type I ER	0.0023	0.0046
lateralization baseline	type II ER	-	-
	type I ER	0.0027	0.0034
lateralization visual stimuli	type II ER	0.816	0.7721
	type I ER	0.0053	0.0115

Note: Type II ERs are not calculated for the baseline conditions as there are no false negatives (there are no local effects expected at any time-space unit).

Table C20

Global ERs by Electrodes Location Value (a priori)

condition	metric	FALSE	TRUE
baseline	type II ER	-	-
	type I ER	0.5949	0.4352
visual left	type II ER	0	0
	type I ER	-	-
visual right	type II ER	0	0
	type I ER	-	-
difference	type II ER	0.2546	0.1782
	type I ER	-	-
lateralization baseline	type II ER	-	-
	type I ER	0.5856	0.3194
lateralization visual stimuli	type II ER	0.0139	0.0093
	type I ER	-	-

Note: Type II ERs are not calculated for the baseline condition as there are no false negatives (no global effect expected), and type I ERs are not calculated for the other tests, as there are no false positives (global effect always expected).

Table C21

Mean FDR by Electrodes Location Value (a priori)

condition	FALSE	TRUE
baseline	-	-
visual left	0.5327	0.0197
visual right	0.5653	0.0917
difference	0.6364	0.2561
lat. baseline	-	-
lat. visual stimuli	0.1401	0.0467

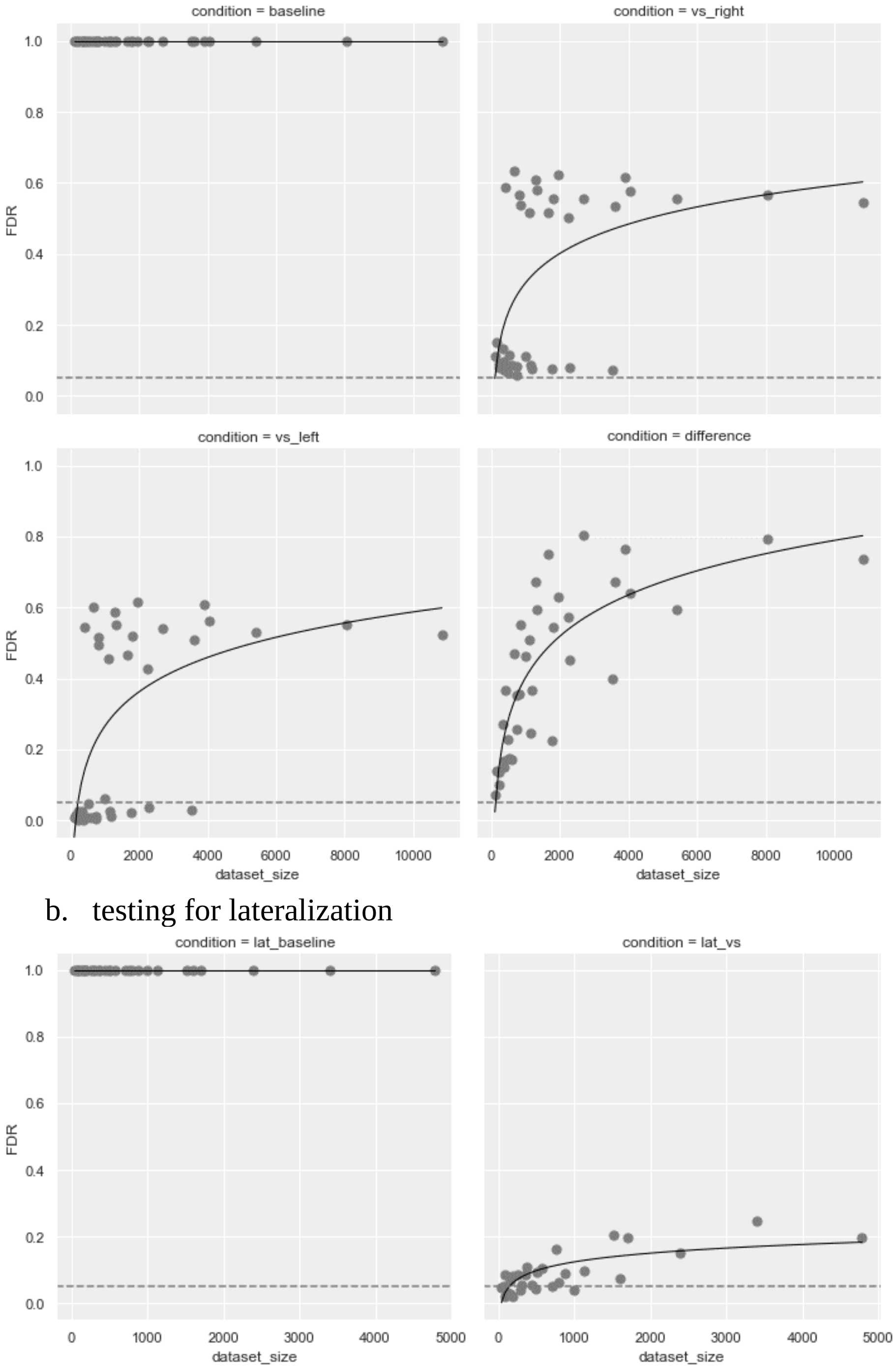
Note: Statistics are calculated based on values from globally significant datasets. FDR is not calculated for the baseline condition, as any positive is a false positive (M = 1, SD = 0).

Appendix D

Performance Metrics Across Datasets

Figure D1

Mean FDR per dataset size



a. testing for signal

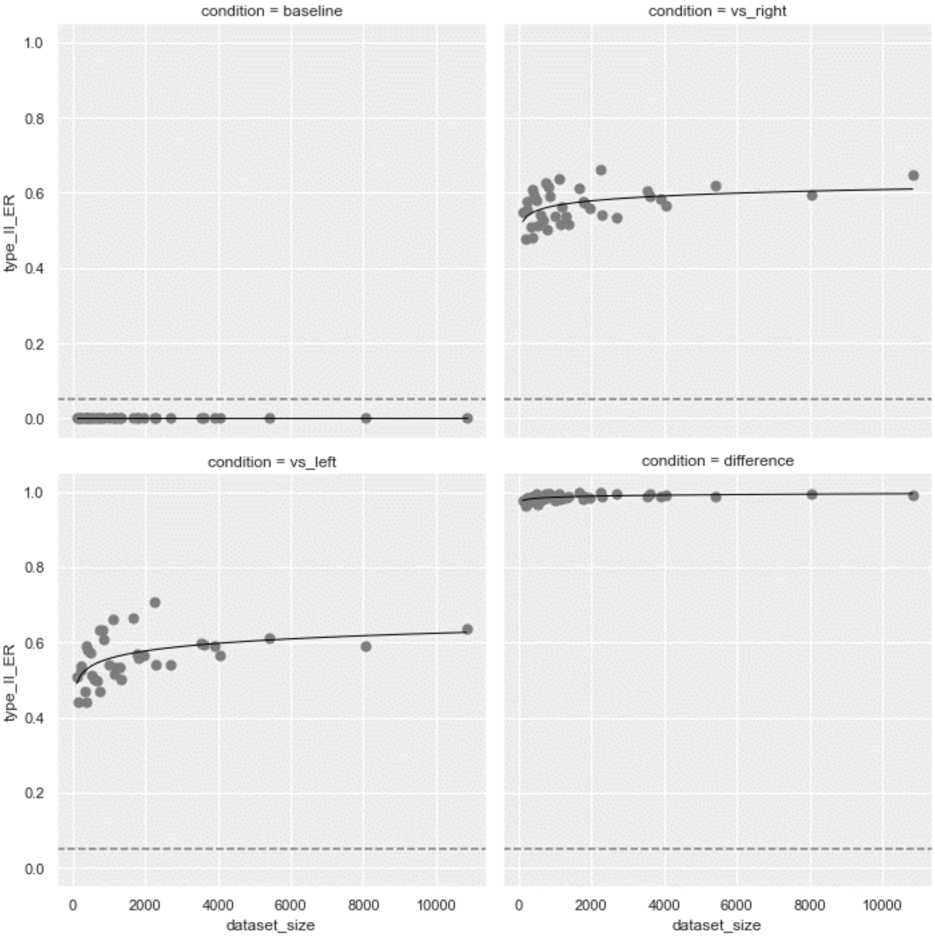
b. testing for lateralization

Note: Dataset size is measured in the total number of time-space units. The average FDR is calculated for every dataset size. FDR increases with dataset size. The scatter plot has been fitted with a logarithmic regression function.

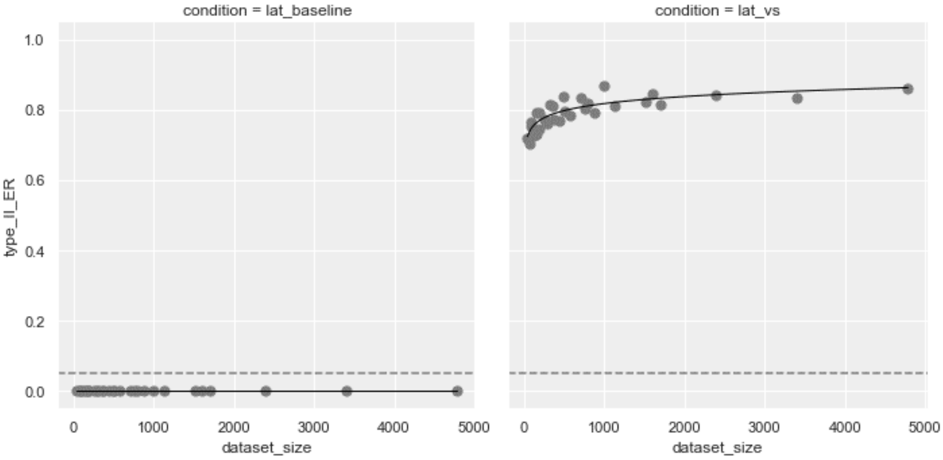
Figure D2

Mean Local Type II ER per dataset size

a. testing for signal



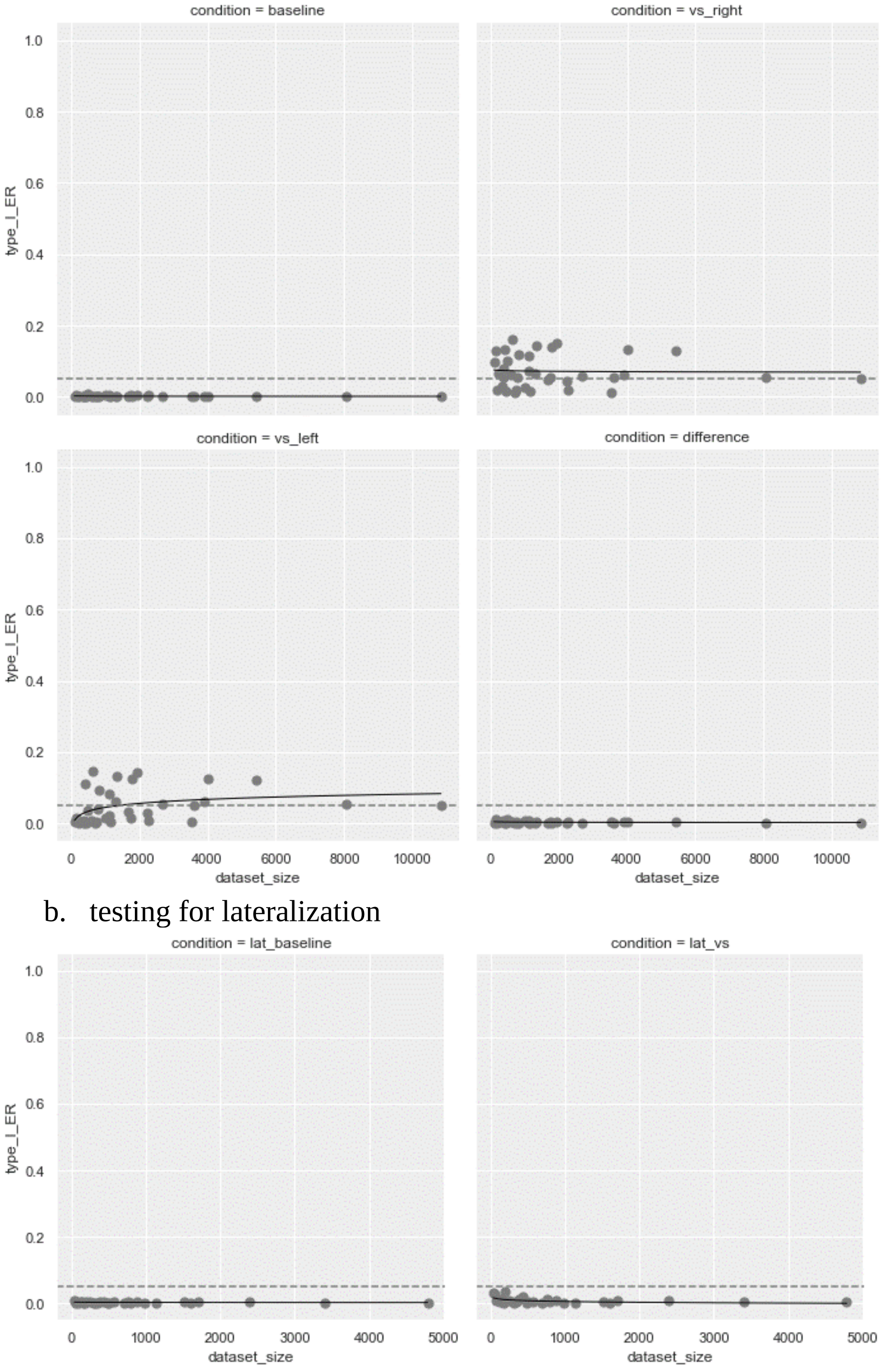
b. testing for lateralization



Note: Dataset size is measured in the total number of time-space units. The average local type II ER is calculated for every dataset size. Type II ER does not seem to be strongly correlated dataset size. Very small datasets (<1000 units) have up to 10% lower type II ER. The scatter plot has been fitted with a logarithmic regression function.

Figure D3

Mean Local Type I ER per dataset size



a. testing for signal

b. testing for lateralization

Note: Dataset size is measured in the total number of time-space units. The average local type I ER is calculated for every dataset size. Type I ER does not seem to be strongly correlated with dataset size. The scatter plot has been fitted with a logarithmic regression function.

Appendix E

Pseudocode for determining the optimal critical *p*-value for one dataset based on baseline data

```
data ← load(prepared_baseline_data);
critical_p_value ← calculate(0.05, data.nr_windows, data.nr_electrodes);
step ← 0.001;

optimal ← False;

while not optimal:
      results ← analyse(data, critical_p_value);
      metrics ← evaluate(results);
      local_type_I_ER ← metrics.local_type_I_ER;

      if local_type_I_ER is 0:
            optimal ← True;
      else:
            critical_p_value ← critical_p_value – step;

optimal_critical_p_value ← critical_p_value;
```