UNIVERSITY OF BELGRADE
FACULTY OF PHILOSOPHY

Anđela Z. Šoškić

EVALUATING ERP METHODOLOGY AND STATISTICS IN EXPERIMENTS USING N400 AFTER PICTURE STIMULI

Doctoral Dissertation

Belgrade, 2019

УНИВЕРЗИТЕТ У БЕОГРАДУ

ФИЛОЗОФСКИ ФАКУЛТЕТ

Анђела З. Шошкић

ПОРЕЂЕЊЕ МЕТОДОЛОШКИХ И СТАТИСТИЧКИХ ПОСТУПАКА У ИСТРАЖИВАЊИМА ПОТЕНЦИЈАЛА У ВЕЗИ СА ДОГАЂАЈЕМ КОД N400 РЕАКЦИЈЕ НА СЛИКОВНУ СТИМУЛАЦИЈУ

докторска дисертација

Београд, 2019.

Supervisor:

Associate Professor Vanja Ković, PhD
Department of Psychology, Faculty of Philosophy
University of Belgrade

Board members:

Assistant Professor Lazar Tenjović, PhD
Department of Psychology, Faculty of Philosophy
University of Belgrade

Assistant Professor Milica Janković, PHD
School of Electrical Engineering
University of Belgrade

Research Associate Andrej Savić, PhD
School of Electrical Engineering
University of Belgrade

Assistant Professor Emily Kappenman, PhD
Department of Psychology
San Diego State University

Date: ________________

Acknowledgements

While working on this dissertation, I was incredibly fortunate to be surrounded by generous people who took time to think about my research, provide valuable guidance, donate their time and skills to resolve obstacles in my way, or to reassure me and take over other work when I was tired or anxious.

In the first place, I am forever grateful to Prof Vanja Ković, who was far more than a mentor to me over the past four years. She has indebted me with her kindness, support and wisdom in so many ways that I can only hope to learn her incredible way of working with students one day and pay my debt forward, because I doubt I will ever get close to paying it back.

I am also thankful to my dissertation committee members, Professors Emily Kappenman, Lazar Tenjović, Andrej Savić, and Milica Janković, for their insightful comments on my dissertation proposal. Professors Tenjović and Savić provided their invaluable advice whenever I had questions. Prof Kappenman gave thorough and thoughtful feedback in various stages of progress of this project, and she was always willing to answer the questions and dilemmas I encountered on this journey.

Prof Guillaume Thierry not only selflessly shared the dataset which was used in Study 2 and provided additional information when I had questions, but also hosted me in his laboratory and sat with me as if I was his own student to show me how he does ERP research. Prof Suzy Styles encouraged my research with great enthusiasm, commented on parts of the text and came up with inspiring suggestions every time we had a chance to meet. A lot of what I learned about ERPs came from the ERP Boot Camp lead by Professors Steven Luck and Emily Kappenman. During the Boot Camp, I also had a valuable opportunity to discuss an early version of my dissertation proposal with Prof Luck.

Vojislav Jovanović set aside an incredible amount of his time to be a meticulous second pair of eyes in Study 1. Katarina Stekić spent hours pre-processing ERP data with me. Božidar Šimšić, Pavle Šoškić, and Lazar Radosavljević saved me a lot of trouble by writing parts of MATLAB code for Study 2. Without Marija Brković and Nemanja Antonijević, Figure 2.2. would probably be a lousy Paint drawing. Thanks are also due to Faculty of Philosophy administrative staff, who helped me navigate the bureaucratic side of my PhD studies, and especially to Maja Jovanović, who was always there when I needed information and advice. My colleagues at the Teacher Education Faculty and fellow PhD students, Anđela Đuković and Ivana Mrvoš, were there for me when I needed advice or help with juggling work and dissertation. I would also like to thank Professors Aurel Božin, Dejan Lalović, and Ksenija Krstić, with whom I worked as Teaching Assistant, for their understanding and encouragement.

I am indebted to tax-paying citizens of the Republic of Serbia, who covered almost all of my tuition costs and additionally supported my research through the Ministry of Education, Science, and Technological Development grant OI179033, as well as the Scholarship for exceptionally gifted students I received in the first year of my PhD studies.

I am thankful to dr Dragan Mijović for introducing me to the work Prof Ković does and for his life-changing advice to contact her to enquire about starting PhD studies with her as my mentor.

I extend my appreciation to friends and family members who provided emotional support, offered warm words of encouragement and had patience for the fact that the time I devoted to this project meant that I had less time and energy for them. In the first place, Lazar took over a lion's share of the burden of our domestic life, and still managed to be nothing but supportive and understanding.

Finally, I am using this opportunity to thank those who set me on the path which has led me here and supported me along the way: my parents, Zlatan and Jelena, grandmother Zenifa, my aunt Elena, and the small army of teachers who have taught me over the past two decades, especially my beloved professors from Kraljevo Grammar School.

Evaluating ERP methodology and statistics in experiments using N400 after picture stimuli

Summary

The knowledge about event-related potentials (ERP) methodology that has accumulated across decades provides useful guidance to researchers on how to make decisions they encounter in ERP experiments. However, while basic standards outline what is not acceptable, there are still many decisions to make when recording and analyzing ERP data, and for each of them, multiple options are acceptable, resulting in thousands of potential unique data pre-processing and analysis strategies.

Having this in mind, the goal of the studies presented here was to explore the way ERP research is done and presented, to examine the consequences of these decisions, to highlight some common issues, and to contribute to advocating for more rigorous methodology and more comprehensive reporting.

These issues were explored in two studies. In Study 1, methodology trends in ERP research were examined using a systematic review approach. In Study 2, data from an existing, published study was used to assess how the variability in basic processing and analysis decisions which is found in the existing literature could influence experimental effects. Due to the diversity of ERP methodology, we focused on a narrow category of ERP studies, those investigating a well-established component (the N400), in the most commonly assessed population (healthy neurotypical adults), in one of its common modalities (visual images).

The studies demonstrate that improvements in reporting on ERP methods are necessary and show which points are most often overlooked. Furthermore, we identify the most common deviations from the guidelines for good practice, as well as methodological decisions which could have influence on ERP effects and statistical power to detect them. Finally, we provide basis for a priori analysis strategies for future studies.

Key words:

ERP methodology, statistics, N400, visually evoked potentials, event related potentials, open science, reproducibility

Scientific field: psychology

Scientific subfield: general psychology

UDC: 159.9.018:159.95(043.3)

Поређење методолошких и статистичких поступака у истраживањима потенцијала у вези са догађајем код N400 реакције на сликовну стимулацију

Апстракт

Сазнања о методологији истраживања потенцијала у вези са догађајем (ERP) акумулирана током деценија пружају корисне смернице истраживачима при доношењу истраживачких одлука. Међутим, иако постоје основни стандарди о томе које праксе нису прихватљиве, истраживачима и даље остаје много одлука које се морају донети приликом снимања и анализе ERP сигнала, и за сваку од њих постоји више решења. Последично, у сваком експерименту постоје хиљаде могућих стратегија обраде и анализе података у сваком истраживању.

Стога, циљ овде изложених истраживања био је да се истраже како начин на који се ERP истраживања спроводе и о њима извештава, тако и последице одлука које се доносе, да се истакну чести проблеми, и да се допринесе настојањима да се истраживања спроводе темељније и о њима извештава детаљније.

Ови проблеми сагледани су кроз два истраживања. У првом су испитани методолошки трендови у области ERP сигнала путем систематског прегледа литературе. У другом су подаци из постојеће, публиковане студије искоришћени да се провери да ли постојећа варијабилност у основним одлукама о обради и анализи ERP сигнала може да утиче на исходе истраживања. Због разноврсности саме ERP технике, умерили смо се на једну, често проучавану меру (N400), добијену у најчешће испитиваној популацији (здрави неуротипични одрасли испитаници), у једном од честих модалитета у којима се она испитује (сликовни стимулуси).

Истраживања су показала да је неопходно побољшати извештавање о методолошким одлукама и омогућила да се укаже на најчешће превиде. Такође, идентификовали смо најчешћа одступања од препорука за добру праксу и неке од методолошких одлука које могу утицати на величину ERP ефеката и статистичку снагу да се они детектују. Коначно, предложили смо основу за доношење а приори одлука о статистичкој анализи ERP сигнала у будућим истраживањима.

Кључне речи:

ERP методологија, статистика, N400, визуелни евоцирани потенцијали, потенцијали у вези са догађајем, отворена наука, репродуцибилност

Научна област: психологија

Ужа научна област: општа психологија

УДК: 159.9.018:159.95(043.3)

Contents

01 Introduction

Event-related potentials, or ERPs, which will be in the main focus of this dissertation, are brief changes in voltage which result from the neural electrical activity related to an event of interest (usually a stimulus or a motor response) (Picton et al., 2000). In the field of neurocognitive research, ERPs are typically the brain's electrical responses to stimuli (evoked potentials), although activity related to other events is sometimes also studied (e.g. preparation for a motor response or processes that follow it).

According to Luck (2014), ERPs were most likely first recorded in 1939 by Pauline and Hallowell Davis, who were investigating differences in the activity of the brain during wakefulness and sleep (H. Davis, Davis, Loomis, Hervey, & Hobart, 1939; P. A. Davis, 1939). Since those early days, ERP analysis has become a method of choice to answer a variety of questions about normal and pathological functioning of the human brain. The number of papers accumulated over decades is huge – for example, just a search for the exact phrase "event related potential*" on the Web of Science gave 26,047 results (on November 05, 2019), in fields ranging from psychiatry, immunology or even obstetrics, to psycholinguistics and educational psychology.

ERPs from recording to results

Brain ERPs are extracted from electroencephalographic (EEG) recordings. In EEG recording, a set of electrodes is placed on the human scalp, and changes in voltage over time are recorded, amplified thousands of times to reveal the minuscule fluctuations caused by the brain's electrical activity, before being digitized and stored. These fluctuations are composed of a series of overlapping waves resulting from a variety of different processes. To extract ERP information from the EEG, timing of ERP-triggering events is recorded together with the electrical brain activity, and the time segments preceding or following the triggers are taken for analysis.

In this section, a description of ERP recording, processing and analysis will be given, and terminology that will be used later in this dissertation will be introduced.

Recording

An EEG recording system consists of several elements. First, it always includes a set of *electrodes*, which are usually placed in a cap for easier placement and positioning. Electrodes positions are described using one of broadly accepted conventions (e.g. Oostenveld & Praamstra, 2001) in order to allow replication and comparison of results. Given that voltage always reflects the difference between two points, there are usually three main types of electrodes. *Active* electrodes[1] are the ones that are placed on locations from which voltages are measured, while *reference electrode* and *ground electrode* are used to create a common reference point for all active electrodes. Voltages on all electrodes (active and reference) are recorded relative to the ground electrode. However, due to the recording system design, the ground electrode necessarily picks up noise from other parts of the system. To eliminate this noise, voltage recorded on the reference electrode is subtracted from active electrodes. This way, noise from the ground electrode is subtracted away, because both recordings from active and reference electrodes include it, and all voltages of active electrodes are expressed relative to the reference electrode. In most studies, there are multiple active electrodes. There is typically one reference electrode, but it can sometimes be split into multiple ends which are placed on different locations to form a *linked reference* (Miller, Lutzenberger, & Elbert, 1991).[2]

To ensure a stable connection between the electrodes and scalp, a conductive medium, usually a gel or saline water, is placed on the contact point. Electrode-skin *impedances* are commonly used as a measure of the quality of this connection. High impedances and high variability of impedances lead to two major problems (Luck, 2014). The first problem is that difference in quality of electrode-scalp connection between an active and reference location can result in less successful elimination of noise picked up by the ground electrode. The second issue is that recording EEG data with high impedances produces skin potentials – slow artifactual changes in voltage (Picton & Hillyard, 1972). These problems are usually eliminated by reducing impedances below a set threshold by gently abrading skin at the points of contact (ideally 2-3 kΩ, but more commonly 5 kΩ). This way, variability in impedances is reduced at the same time. Some recording systems, called high input impedance systems, are less sensitive to the first problem, but skin potentials remain an issue in any case, especially when low-frequency components, such as the N400, are measured (Kappenman & Luck, 2010).

Signal registered by electrodes is transmitted to amplifier, where it is amplified and filtered before being digitized. Filters remove voltage changes whose frequencies are outside the range of EEG activity, like slow shifts due to changes in electrode-scalp impedances over time and 50/60 Hz power-line noise. *High-pass* filters remove low frequencies, *low-pass* filters remove high frequencies, *band-pass* filters leave filters within a certain range, while *band-stop* filters remove a range of frequencies. *Notch* filters are band-stop filters with a narrow stopband. Describing filters requires providing several parameters, because filters typically do not have a clear cut-off point beyond which all frequencies are eliminated. Instead, there is a transition band – the range within which frequencies are eliminated more

[1] Alternatively, term "active electrode" is sometimes also used to refer to a specific kind of electrodes, which are made with built-in pre-amplifiers to reduce external noise during recording.

[2] Some authors use term „linked reference" to refer to offline averaging of multiple reference sites. Following recommendations by Luck (2014), to avoid ambiguity, this term will only be used to refer to physically linked electrodes to avoid ambiguity.

and more. ERP guidelines (Keil et al., 2014) recommend reporting filter type, half-amplitude or half-power cut-off point and its frequency response function slope at the cut-off point (for a discussion about these properties, see Cook & Miller, 1992).

Data preprocessing

After being recorded, EEG data is subjected to a series of procedures collectively called data preprocessing, whose goal is to prepare data for ERP analysis and eliminate noise that cannot be sufficiently eliminated using averaging (Keil et al., 2014).

Digital filtering.

In addition to being filtered online (*analog filtering*), EEG data is sometimes also filtered offline (*digitally*). Because analog filtering cannot be reversed and digital filters have advantages over analog ones, it is usually recommended that data is filtered online as little as possible, and that the rest of the filtering is carried out offline (Luck, 2014).

Artifact rejection and correction.

Raw EEG recordings contain artifacts - noise from other sources in addition to brain activity. Some of these artifacts, such as 50/60 Hz line noise from the environment, can be removed by filtering. Some can be small and infrequent enough that the researcher can count on trial averaging to eliminate them (this is often the case with EKG). However, there are artifacts, primarily eye blinks, but also eye movements and others, which can remain after averaging. These artifacts are usually removed by excluding trials contaminated with artifacts (*artifact rejection*) or by applying procedures that estimate the shape of artifacts and subtract them from the EEG, leaving “clean” data (*artifact correction*). In some studies, artifact correction and rejection are combined. Most artifact correction procedures require that large artifacts be removed before artifact correction to avoid contaminating artifact models created by correction algorithms (Luck, 2014). Additionally, artifact rejection can be used to eliminate residual artifacts that cannot be removed using correction procedures.

Segmenting.

As mentioned earlier, ERPs are extracted from EEG by taking segments of data time-locked to the events of interest. These segments are called *epochs*, and in case of evoked potentials, they typically include a short prestimulus period and a poststimulus period long enough to include all components of interest.

Baseline correction.

Despite applying high-pass filters, some slow changes in voltage always remain, and individual trials, and even averaged trials, are not necessarily aligned on y-axis. This is remedied by *baseline correction*. In case of evoked potentials, baseline correction is usually conducted by subtracting the average of prestimulus period from each time point in epoch, because it does not contain any stimulus-elicited activity. Epoch prestimulus period and baseline period are usually the same and many ERP data processing systems automatically perform baseline correction when segmenting data into epochs, but this is not necessarily the case.

Re-referencing.

Voltages reference point can be *re-referenced,* or changed offline to a different reference electrode, or even an average or sum of multiple reference electrodes. Multiple reference sites are often used to obtain a reference which is hemisphere-neutral, usually the average of electrodes placed on earlobes or mastoid processes, or the average of all electrode sites (*average reference*).

Averaging

While the EEG fluctuations are small and they can only be seen on the scale of dozens of microvolts, ERPs are even smaller (effects can sometimes be smaller than 1 µV). Namely, ERPs represent only a small portion of the overall EEG, hidden among the many unrelated brain processes taking place at the same time. To extract this information, ERP analysis requires recording many responses, which are typically averaged together to eliminate the noise and to preserve only the changes that are systematically related to the events of interest.

Measurement and statistical analysis

In most studies, ERPs are examined in the time domain, as ERP waveforms. ERP waveforms are short segments of EEG data, which show changes in voltage over time, and consist of a series of positive and negative deflections, called *waves* or *peaks*. Averaging epochs produces separate waveforms for each participant, each experimental condition and each electrode site, or sometimes a group of electrode sites, called *region of interest* (ROI).

The shape of each waveform is a result of superimposition of multiple underlying *components*, which represent separate neural processes that are part of an ERP response (Picton & Stuss, 1980). Even in simple experimental tasks, ERPs necessarily comprise a variety of processes, which are part of stimuli perception, attention, categorization, language processing, and many other functions that can be engaged during task completion. To overcome the problem of analyzing and interpreting such results, ERP researchers have developed dozens of experimental paradigms and studied the effects of experimental manipulations on the ERP waveform. As a result, a variety of ERP waves have been isolated and described in terms of their temporal and spatial distribution, tasks that elicit them and their sensitivity to experimental manipulations, their proposed meaning and, in some cases, their source in the brain. These waves are considered ERP components, although some of them likely represent a composite of multiple simultaneous subcomponents, which have not been separated yet.[3]

[3] To highlight this difference between what components are in theory and the changes in the ERP waveform which researchers call components, Luck (2014) makes the distinction between the conceptual and operational definition of the ERP component. In theory, "an ERP component is a

A typical ERP study, therefore, is designed to utilize properties of known components to allow interpretation of the results. ERP variables in these studies are measures that describe individual waves or differences between two or more waveforms, e.g. mean amplitude over a time interval, or latency at which two waveforms start to diverge. Additionally, because ERPs are recorded at different scalp locations, spatial distributions of the experimental effect can be compared. Measures obtained this way can easily fit as dependent variables in traditional experimental designs, in which they are often treated in a similar way to behavioral variables such as response time (RT) or accuracy (for a more detailed review on how ERPs can be used in research, see Kappenman & Luck, 2011; Luck, 2014).

Order of operations

The above described preprocessing operations, averaging and measurement can be implemented in a different order. Because not all steps are linear, the outcome may vary depending on the order of operations, and for some steps, one sequence is preferable to another (Luck, 2014). For example, it is recommended that filtering, especially high-pass filtering, is applied to continuous data, because filtering produces artifacts at the beginning and end of the data segment to which it is applied. As described, it is desirable to eliminate large artifacts prior to artifact rejection, which is conducted on continuous data. However, if artifact rejection is used to eliminate blinks and other smaller artifacts and it is conducted automatically using an absolute voltage threshold, this step should be applied after epoching and baseline correction. For this reason, it is important that the order of operations is carefully considered, and clearly outlined together with other information on ERP methods (Keil et al., 2014).

The N400: an ERP component example

As we will see later in this chapter, the studies that are presented in this dissertation focus on ERP methodology and data analysis. However, ERP analysis not only very versatile, but also highly diverse. An ERP study's methods, and processing and analysis pathways depend to some extent on the study design: for example, on components that are being measured, the modality of the stimuli, and the population from which subjects are recruited.

Studying ERP methodology in general would require taking all these factors into account, in addition to the methodology parameters that are being examined. To overcome this issue, we chose to focus on a narrow category of ERP studies, those investigating a well-established component (the N400) in the most commonly assessed population (healthy neurotypical adults), in one of its common modalities (visual images). N400 is a well-known

scalp-recorded neural signal that is generated in a specific neuroanatomical module when a specific computational operation is performed" (Luck, 2014, p. 66). Operationally, an ERP component is "a set of voltage changes that are consistent with a single neural generator site and that systematically vary in amplitude across conditions, time, individuals, and so forth. That is, an ERP component is a source of systematic and reliable variability in an ERP data set" (Luck, 2014, p. 68).

ERP component with a long history of successful conceptual replications (Kutas & Federmeier, 2011), making it an ideal target for investigations of methodological and analytical problems in the field. The findings of the presented research are directly relevant to a large group of N400 researchers, while some of the points are of relevance to the entire ERP field. It is also a large component, and as a result more robust to changes in analysis procedures which are examined in our Study 2. If a robust component turns out to be sensitive to variations in data processing and analysis pipeline, one should be even more careful with smaller components, which are easier to obscure with noise or to incorrectly register.

In the following section, some basic properties of the N400 will be described. (More detailed accounts can be found in reviews by Kutas & Federmeier (2011) and Swaab, Ledoux, Camblin, & Boudewyn (2012).)

N400 (also: N4, usually in older papers) is by far the most extensively studied language-related ERP component, and one of the most researched components in general. Over decades, the N400 has been used as a variable in thousands of studies (e.g. search for "N400" on Web of Science gave 2,983 returns (on November 05, 2019)), primarily focusing on language and semantic memory.

When it was first registered, almost four decades ago, it was the first ERP component that was found to be sensitive to abstract stimuli properties (Kutas & Hillyard, 1980). In this study, Kutas and Hillyard famously observed a negative-going wave peaking at about 400 ms after onset of a semantically unexpected sentence ending (e.g. "He spread the warm bread with socks" vs. "He spread the warm bread with butter"), which was not the case when the final word was semantically expected, but physically different than the words presented before it.

Studies that followed have demonstrated that the factor which produces the N400 effect is not violation of a learned sequence, sentence meaning or truth value, but low probability of the stimulus given its semantic context (Kutas & Federmeier, 2011) , which produces an increase in N400 amplitude. For example, a sentence can be meaningful, and even true, but if its ending is unexpected, it will produce a larger N400. Today we know that the N400 is part of the neural response to words presented in all modalities (written, spoken, sign language), as well as other meaningful stimuli (actions, drawings, mathematical operations) (Kutas & Federmeier, 2011). Sentences are not the only context that can produce the N400 effect – it can be observed, for example, in the semantic priming paradigm (Bentin, McCarthy, & Wood, 1985), or in response to visual images that contain objects that do not fit the rest of the picture (Ganis & Kutas, 2003).

Unlike N400 amplitude, its latency is relatively stable. The component occurs between 200–600 ms poststimulus (Kutas & Federmeier, 2011), and reaches its maximum between 380–440 ms in young neurotypical population (Swaab et al., 2012). With aging, both latency and amplitude decline linearly (Iragui, Kutas, & Salmon, 1996). Regarding spatial distribution, the component has a centroparietal location, when measured relative to a posterior reference such as mean mastoids or earlobes, although the distribution partly depends on the stimulus type that is used (Kutas & Federmeier, 2011; Swaab et al., 2012). Notably, picture-evoked N400 is more frontally distributed than the effect produced by verbal stimuli (Ganis, Kutas, & Sereno, 1996). The differences found in scalp distribution have sparked discussion about whether there are different subtypes of the N400, and, more broadly, attempts have been made to localize the N400 source. The results point out to the N400 being activity of a dynamic system, rather than a single neural source, which is modality-dependent, but not modality-specific (Kutas & Federmeier, 2011).

ERP analysis complexity and the garden of forking paths

ERP data recording, preprocessing and analysis is an incredibly complex process, and researchers are required to make and balance numerous decisions when planning their study and recording data, and even more choices are made on the way from raw EEG recordings to the results of statistical analyses. Of course, the researchers do not have to start from scratch when making these decisions. Parallel with progress in learning about ERP components and using them to study neuropsychological phenomena, papers on ERP methodology have provided answers and guidance to ERP research community (e.g. Boudewyn, Luck, Farrens, & Kappenman, 2018; Delorme, Sejnowski, & Makeig, 2007; Junghöfer, Elbert, Tucker, & Braun, 1999; Kappenman & Luck, 2010; Miller et al., 1991; Tanner, Morgan-Short, & Luck, 2015).

Moreover, as the popularity of the method and its availability to laboratories across the world grew with time, the need for clear and widely available practice guidelines and standards increased. In response, the first guidelines for ERP recording were published in 1977, derived from the International Symposium on Cerebral Evoked Potentials in Man held in Brussels in 1974 (Donchin et al., 1977). These guidelines were updated by the Society for Psychophysiological Research, which published a much more comprehensive document in 2000 (Picton et al., 2000), and the most recent revision came in 2014, in a broader report which focused on electroencephalography as well as magnetoencephalography (Keil et al., 2014). In addition, specialized guidelines have been developed for fields that require a distinct approach, such as clinical studies (Duncan et al., 2009; Kappenman & Luck, 2016) or experiments with children (Taylor & Baldeweg, 2002). Additionally, methodology books on ERP have also been published to help new researchers get acquainted with the basics and provide a more thorough overview (Handy, 2005; Luck, 2005, 2014).

The knowledge about ERP methodology that has accumulated across decades provides useful guidance to researchers on how to make decisions they encounter in ERP experiments. However, while basic standards outline what is *not acceptable*, there are still many decisions to make when recording and analyzing ERP data, and for each of them, multiple options are acceptable. This necessarily puts a researcher in a dilemma over which way to go and opens a possibility of intentional and unintentional data manipulation in order to fit results to expectations.

An example of this issue in the context of ERP research is described in a recent paper by Luck and Gaspelin (2017), who demonstrated how selecting the time window and electrode set for statistical comparison based on visual inspection is equivalent to post hoc selection of one of an almost unlimited variety of possible statistical comparisons, and, consequently, in the probability of a false significant finding approaching certainty.

This problem is not unique to ERP analysis – on contrary, it has been recognized in other fields, as well. For example, one review of methods reporting in fMRI (Carp, 2012) has shown that there are almost as many analyses pipelines for fMRI data as there are individual studies, and many papers fail to provide sufficient information on methods to allow precise independent replications. More broadly, terms *researcher degrees of freedom* (Simmons, Nelson, & Simonsohn, 2011) and *garden of forking paths* (Gelman & Loken, 2013), have been coined to describe the problem of inadequate treatment of post hoc methodological decisions, which leads to an increase in Type I error rate. As Gelman & Loken (2013) explain, almost any dataset can be analyzed and presented in many ways, and even seemingly unambiguous hypotheses can have multiple operationalizations. Many of these methodological decisions are made only after inspecting the data, leading to data-dependent decisions. This problem is particularly of concern in studies involving abundance of data that can be treated in a multitude of ways, which is characteristic of ERP research.

These issues are not just a theoretical concern, as it has been demonstrated recently when a large collaborative preregistered replication attempt (Nieuwland et al., 2018) failed to support the key findings of an influential, widely cited study on the N400 in response to articles and nouns (DeLong, Urbach, & Kutas, 2005). In addition to highlighting the importance of careful design of new studies and replication attempts, obtaining reliable effects and transparent reporting, the study by Nieuwland et al. (Nieuwland et al., 2018)nand ensuing commentaries (DeLong, Urbach, & Kutas, 2017; Yan, Kuperberg, & Jaeger, 2017) provide further evidence of the sensitivity of ERP analysis. Namely, Nieuwland et al. (2018) report that one of the issues raised after publishing a preprint of their paper was the difference in baseline duration between the original study by DeLong et al. (2005) and their replication attempt. The discrepancy in methods section resulted from omission of baseline information from the paper by DeLong et al., and it was corrected after communication between the two author teams following preprint publication. This discussion about baseline differences demonstrates that ERP data processing and analysis choices are not trivial and that comprehensive reporting on these choices is important.

Present study

Given the variability of methodological options, and the potential for processing decisions to influence study outcomes, it is important to understand how ERP studies are conducted in practice, and how variability found in the existing literature affects experimental outcomes. Additionally, it is not currently known to what extent researchers are transparent about their data collection and analysis procedures.

These issues were explored in two studies, and their results are presented in this dissertation. In the first study, methodology trends in ERP research were examined using a systematic review approach, to discover the limits within which ERP data recording, processing, and analysis decisions vary in the peer-reviewed literature. In the next phase, data from an existing, published study were used to assess how the variability in basic processing and analysis decisions which is found in the existing literature could influence experimental effects. As it has been described earlier, due to the diversity of ERP methodology, the two studies will focus on one chosen subgroup of ERP studies: picture-evoked N400 experiments, conducted with healthy neurotypical adults as participants.

Study 1: systematic review of ERP methodology

The aim of this study was to provide a comprehensive overview of the present state in the field, as a platform from which to develop guidance for future neuropsychological research. The questions of interest were: (1) how much methodological variability exists among studies that would be expected to follow similar procedures, because they all investigate the same well-established neurological phenomenon; (2) which practices are the most prevalent; (3) how often researchers deviate from guidelines for good practice; (4) which deviations are the most common; (5) how often descriptions of methods and analyses are insufficiently detailed, and (6) which are the principal areas where improvements in reporting practices are necessary. Answering these questions allowed us to: (1) provide evidence-based guidelines for making decisions about the analysis pipeline, for example, when a priori decisions are made based on previous research (e.g. choosing a reference site or the N400 time window), (2) caution researchers against the most common deviations from best practice in ERP methodology and ERP reporting, and (3) choose variations in data processing/analysis pipeline which were used in Study 2. In order to provide the most robust dataset from which to draw conclusions, we conducted this survey of the existing literature in the form of a systematic review.

The review provides an extensive insight into a variety of parameters, including properties of study design (e.g. sample size), data processing (e.g. filtering procedures), measurement (e.g. N400 time window), statistics (e.g. electrode sites in the ANOVA model), and, for more recent papers, references to supplemental information (e.g. raw data or analysis codes).

Study 2: effects of varying processing and analysis steps in an N400 experiment

As it will be shown in Chapter 2, there was little consistency in the way ERP data was processed and analyzed in the studies included in the systematic review. Similar to the findings in the fMRI field (Carp, 2012), there were almost as many different approaches to ERP data analysis as there were papers, and there were few points on which the majority of researchers took the same approach.

Given this variability, an important question is: does it matter? In other words, would the conclusions of an experiment be the same regardless of the processing and analysis pathway, as long as the processing decisions are within the limits that can be found in the peer-reviewed literature?

To answer this question, we tested the effects of variations in processing steps that can be found in the existing literature on the results of one sample experiment. A sample of data was taken from a published, peer-reviewed N400 study, conducted by an independent team - Boutonnet, McClain, & Thierry (2014) . Data from an already published study was chosen for two reasons. First, the study's design and data quality are already verified by the scientific community. Second, having one known and accepted combination of processing parameters allowed using them as a standard for comparing outcomes of other analyses.

A set of variations in data preparation and statistical analysis was defined based on the systematic review, and these different options were applied to the dataset collected by Boutonnet et al. (2014). The factors that were varied include characteristics of all steps in the processing and analysis pipeline: offline filter high-pass and low-pass cut-offs, eye artifact elimination method (correction vs. rejection), baseline correction time window, N400 amplitude measure, N400 measurement window, reference site(s), and analysis montage. To reduce the number of possible combinations, each parameter was varied independently, while the other decisions were kept the same as in the study by Boutonnet et al. (2014).

Conclusions from different analyses of the same data were compared to see which steps make a difference in study outcomes or lead to a statistically different N400 effect, and which steps are of less concern. Furthermore, Monte Carlo simulations were performed to determine how each of the parameters influences statistical power to detect experimental effects.

* * *

In the chapters that follow, the results of the systematic review are presented first (Chapter 2), followed by a report on the analyses conducted in the second study (Chapter 3). Finally, Chapter 4 contains a general discussion of the results of both studies, a summary of recommendations for conducting research and reporting on it which followed from the present study, and final thoughts with suggestions for future research.

02 Systematic review of N400 ERP methodology[4]

How much variability is there in the peer-reviewed literature?

Objective

This systematic review examines the diversity of methodologies used, and clarity of reporting in peer-reviewed ERP papers reporting an N400 to a visual image stimulus and recorded in adult healthy participants, published between January 1980 – June 2018 in journals included in two large databases: Web of Science and PubMed.

Method

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, Altman, & Group, 2009) where it was applicable. The PRISMA checklist for our review is available in the online repository for this article (Appendix B).

Database Search

The first step was to search online databases for papers relevant for this review. Two large aggregated databases were chosen: Web of Science and PubMed. These two databases contain a large sample of ERP studies, which is likely representative for the majority of peer-reviewed ERP literature.

Each database was searched using the following search terms: (*N400* or *ERP N4*) AND (*visual stimuli*, *visually evoked potentials, drawing(s), image(s), photo(graph-ies,y,s)* or *picture(s)*). Default settings for search engines were used on both platforms: search for key words in all fields and automatically generated MeSH terms for PubMed, and search within Topic for the Web of Science. A list of exact search phrases with numbers of hits for each conducted search is available in Appendix C. Search was limited to papers published after 1980, the year of the N400 discovery. It took place on 11th July 2018, and included papers published until 30th June 2018.

All references were merged into a single database using Mendeley Desktop (Mendeley Ltd.) to identify duplicate publications from the two sources.

Article Scanning

Following the PRISMA procedure, in order to identify which of the unique articles returned by the search did indeed contain a N400 study relevant for our review, we screened each article for possible inclusion. Two researchers independently conducted the screening,

[4] Online supplements to this study are listed in Appendix A.

and where ambiguity or disagreement between the independent screeners arose, additional team members asked to clarify or expand the initial criteria for eliminating studies.

The main criterion for selection was that the papers were original research papers on studies that included an ERP experiment with images as stimuli, and where N400 after the images was explored. Studies which included simultaneous presentation of information in various modalities or rapid presentation of visual image stimuli were not considered due to an effect of such designs on the N400 properties and analysis. For the same reason, papers were excluded if they involved any interventions or recording equipment which could affect experiment methodology or data analysis (e.g. tDCS, fMRI). Studies were selected for analysis only if participants were adults with no reported history of psychopathology.

Contrarywise, there was no limitation regarding methods and outcome measures, since we focused on methodology, and not on results. We also did not exclude 15 studies that involved a task with other types of target stimuli in addition to the task with visual images. Finally, there was no upper limit for participants' age. The N400 is known to change linearly with age (Kutas & Iragui, 1998), so any cut-off point would have been arbitrary, and it was relatively common for studies in our sample to have at least one or two middle-aged participants. Consequently, we included two aging studies with elderly participants. We compared the N400 measurement and analysis montage of aging studies and studies including targets in other modalities, but there were no discrepancies from the overall results.

The review was limited to articles in English, since most papers on ERP are published in this language. On the other hand, studies conducted in other languages, but reported in English, were included in the pool. Additionally, the focus of this review were papers that had been verified and accepted by the scientific community. For this reason, we did not look for papers that were not published or in press at the time of search. Furthermore, we checked papers for retractions and corrections. Conference proceedings were included in the pool if they were full-sized papers, while short resumes or abstracts were excluded due to the typical lack of methodological detail in the short format.

In addition, references were excluded if they could not be located through their journal or web search. Publications were considered duplications and duplicates were excluded if multiple papers had the same study design, sample characteristics, and statistical results. In cases where papers, potentially or expressly, reported different analyses of the same data, all versions were included. Since we focused on methods, these papers added new information to our review, and they overlapped only in study design and pre-processing, which would likely have been the same if the authors had collected new data for each analysis.

Data extraction

All papers were independently assessed by two researchers, who reported the results in separate spreadsheets. The two spreadsheets were then merged, and all diverging or unresolved points were jointly analyzed by one of the authors working on papers assessment and a third team member. When a conclusion about a reported item could not be reached due to conflicting, insufficient or ambiguous information, it was labeled as "inconclusive". In case of some variables, categories could not be made in advance. In these cases, descriptions were logged and merged using the procedure above, and categorization was carried out post hoc by one team member.

Data was extracted for the following properties, using a total of 74 columns (variables):

- *experimental design:* design description, smallest sample size[5] – total and per group, smallest number of trials – total and per situation, jittering pre-stimulus intervals, use of techniques to prevent overlap between the overt response and ERP window;
- *equipment:* hardware used for EEG recording (cap, amplifiers, other), software used during the experiment and data processing and analysis (stimulus presentation, EEG acquisition, EEG/ERP processing, statistics, other);
- *data recording and processing:* reference used in data analyses, recording montage (active sites), scalp electrodes impedance, basic low-pass and high-pass online and offline filter settings (cut-off, roll-off, and cut-off type – half-amplitude or half-power), use of notch filters, number of trials left after data processing – what type of information was reported and what were the numbers, baseline length, epoch duration and whether it overlapped with an overt response or the beginning of the next trial, which artifacts were eliminated, artifacts identification and elimination procedures, whether the order of operations could be at least assumed based on the description;
- *measurement:* N400 time window and the reason for selecting this specific window, amplitude measure;
- *statistical analyses and data presentation*: which electrodes or electrode constellations were analyzed (analysis montage), electrode analysis strategy (basis for choosing analysis montage), main analysis approach (e.g. ANOVA model), additional analyses (e.g. post hoc tests, topographical analyses), whether there was correction for sphericity violation and having multiple statistical tests, number of uncorrected (M)AN(C)OVAs, how many other components were analyzed in addition to N400, which additional components were analyzed and whether they were earlier or later than the N400, whether negative was plotted up or down in the graphs;
- *about publications:* publishing year, authors, whether it was a conference proceeding or a journal article;
- *general*: a column for additional data and comments.

Finally, *availability of supplemental data* (e.g. stimuli, raw data), identifiable through the article, was examined. This is a more recent trend in scientific reporting, and we did not

[5] In some cases, there was more than one experiment in a paper. Furthermore, individual experiments could have uneven groups or an uneven number of trials per condition. In these situations, we chose the lowest number, because we were interested in how often authors deviated from the guidelines for good practice.

expect most papers to provide this information. However, there has been a push in the past few years towards improving reproducibility and credibility of research through encouraging open science practices (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Nosek et al., 2015), so we were interested whether more recent papers had started to implement these recommendations.

Due to the volume of information, variable descriptions and coding details are provided in a Codebook in Appendix D. The results were summarized by examining descriptive statistics: frequencies of categorical variables, as well as means and standard deviations of numerical variables. In rare cases, where it was not possible or rational to categorize papers due to extreme variability, verbal descriptions were summarized by examining frequencies of key words.

Results and discussion

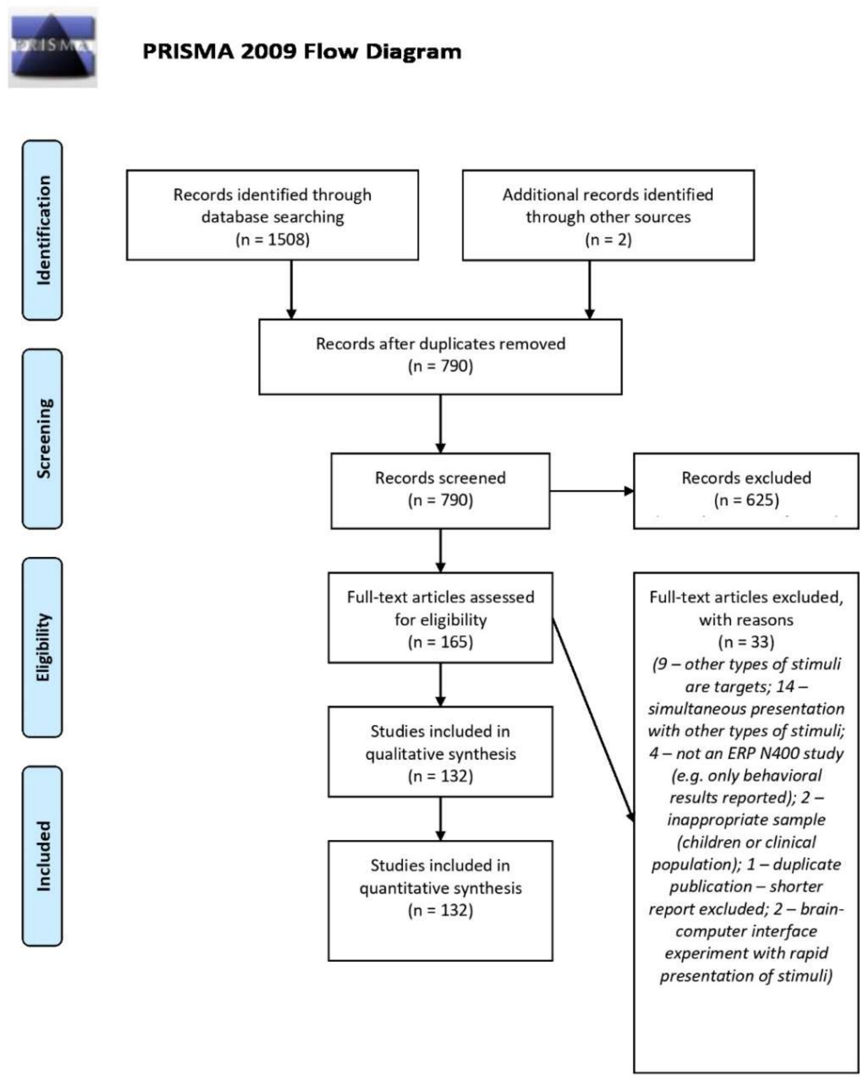
Database search and article selection

In total, 1508 papers were returned by the searches. Two additional references were added, which were found during a preliminary stage of the systematic review, but they did not show up in database search results. After merging search results and removing duplicates, 790 titles remained.

Of these, 625 articles were excluded on inspection of title and abstract, and 33 were excluded after inspecting the full text, leaving 132 papers for inclusion in the review. Out of 790 papers returned by the searches, 17 were in languages other than English. Three references were excluded because they could not be located through their journal or web search. One paper was excluded because it was a duplicate publication. Eighty-three papers did not include an ERP N400 experiment (e.g. theory papers, intracranial recordings), and others were rejected based on their methods (sample or study design).

There was only one correction, no retractions, and it concerned a name spelling error. Six conference proceedings were included in our review, and the remaining papers were journal articles.

The PRISMA flow diagram summarizing articles included/excluded at the different stages of screening can be seen in Figure 2.1. The full list of all papers included in this report can be found in Table 2.1. Online supplements 5 and 6 contain the spreadsheet with extracted information on individual papers and Excel files with all results and graphs presented here (see Appendix A).


PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1508)
Additional records identified through other sources (n = 2)
Records after duplicates removed (n = 790)
Records screened (n = 790)
Records excluded (n = 625)
Full-text articles assessed for eligibility (n = 165)
Full-text articles excluded, with reasons (n = 33) (9 – other types of stimuli are targets; 14 – simultaneous presentation with other types of stimuli; 4 – not an ERP N400 study (e.g. only behavioral results reported); 2 – inappropriate sample (children or clinical population); 1 – duplicate publication – shorter report excluded; 2 – brain-computer interface experiment with rapid presentation of stimuli)
Studies included in qualitative synthesis (n = 132)
Studies included in quantitative synthesis (n = 132)

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.

Figure 2.1. PRISMA flow diagram.

Table 2.1. Papers evaluated in this report, in chronological order by year and alphabetical order within a year

No.	Decade	Study
1	1980s	(Barrett, Rugg, & Perrett, 1988)
2		(Barrett & Rugg, 1989)
3	1990s	(Barrett & Rugg, 1990)
4		(Friedman, 1990)
5		(Nigam, Hoffman, & Simons, 1992)
6		(Stuss, Picton, Cerri, Leech, & Stethem, 1992)
7		(Bobes, Valdes-Sosa, & Olivares, 1994)
8		(Holcomb & McPherson, 1994)
9		(Perez-Abalo, Rodriguez, Bobes, Gutierrez, & Valdes-Sosa, 1994)
10		(Pratarelli, 1994)
11		(Nielsen-Bohlman & Knight, 1995)
12		(Schweinberger, Pfütze, & Sommer, 1995)
13		(Yano, 1995)
14		(Debruille, Pineda, & Renault, 1996)
15		(Ganis et al., 1996)
16		(Pietrowsky et al., 1996)
17		(Simos & Molfese, 1997)
18		(Mecklinger, 1998)
19		(Münte et al., 1998)
20		(Grigor, 1999)
21		(Jordan & Thomas, 1999)
22		(McPherson & Holcomb, 1999)
23		(Olivares, Iglesias, & Bobes, 1999)
24	2000s	(Castle, Van Toller, & Milligan, 2000)
25		(Eimer, 2000)
26		(Kiefer, 2001)
27		(Bensafi et al., 2002)
28		(Federmeier & Kutas, 2002)
29		(Hamm, Johnson, & Kirk, 2002)
30		(West & Holcomb, 2002)
31		(Ganis & Kutas, 2003)
32		(Jemel, Calabria, Delvenne, Crommelinck, & Bruyer, 2003)
33		(Mnatsakanian & Tarkka, 2003)
34		(Olivares, Iglesias, & Rodríguez-Holguín, 2003)
35		(Schendan & Kutas, 2003)
36		(Yuping Wang et al., 2003)
37		(Wicha, Bates, Moreno, & Kutas, 2003)
38		(Wicha, Moreno, & Kutas, 2003)
39		(Gunter & Bach, 2004)
40		(Mnatsakanian & Tarkka, 2004)
41		(Trenner, Schweinberger, Jentzsch, & Sommer, 2004)
42		(Yuping Wang, Cui, Wang, Tian, & Zhang, 2004)
43		(Yovel & Paller, 2004)
44		(Balconi & Pozzoli, 2005)
45		(Gierych, Milner, & Michalski, 2005)
46		(Supp et al., 2005)
47		(Eddy, Schmid, & Holcomb, 2006)

No.	Decade	Study
48		(Paz-Caballero, Cuetos, & Dobarro, 2006)
49		(Cooper, Harvey, Lavidor, & Schweinberger, 2007)
50		(Mao & Wang, 2007)
51		(Proverbio, Del Zotto, & Zani, 2007)
52		(Wu & Coulson, 2007)
53		(Boldini, Algarabel, Ibanez, & Bajo, 2008)
54		(Hirschfeld, Jansma, Bölte, & Zwitserlood, 2008)
55		(Koester & Schiller, 2008)
56		(Lüdtke, Friedrich, De Filippis, & Kaup, 2008)
57		(Neumann & Schweinberger, 2008)
58		(Ortega, Lopez, & Aboitiz, 2008)
59		(Steffensen et al., 2008)
60		(Zhang, Li, & Zhou, 2008)
61		(Eddy & Holcomb, 2009)
62		(J. D. Harris, Cutmore, O'Gorman, Finnigan, & Shum, 2009)
63		(Kovic, Plunkett, & Westermann, 2009)
64		(Proverbio & Riva, 2009)
65		(Shibata, Gyoba, & Suzuki, 2009)
66	2010s	(Eddy & Holcomb, 2010)
67		(Khateb, Pegna, Landis, Mouthon, & Annoni, 2010)
68		(Liu et al., 2010)
69		(Lu et al., 2010)
70		(Mudrik, Lamy, & Deouell, 2010)
71		(Olivares & Iglesias, 2010)
72		(Proverbio, Riva, & Zani, 2010)
73		(Saavedra, Iglesias, & Olivares, 2010)
74		(Eddy & Holcomb, 2011)
75		(Herring, Taylor, White, & Crites, 2011)
76		(Huffmeijer, Tops, Alink, Bakermans-Kranenburg, & van Ijzendoorn, 2011)
77		(Kiefer, Sim, Helbig, & Graf, 2011)
78		(Kuipers & Thierry, 2011)
79		(Liao, Su, Wu, & Qiu, 2011)
80		(Lin, Wang, Cheng, & Cheng, 2011)
81		(Maillard et al., 2011)
82		(Wu & Coulson, 2011)
83		(Yum, Holcomb, & Grainger, 2011)
84		(Blackford, Holcomb, Grainger, & Kuperberg, 2012)
85		(Bramão et al., 2012)
86		(Cansino, Hernández-Ramos, & Trejo-Morales, 2012)
87		(Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012)
88		(Demiral, Malcolm, & Henderson, 2012)
89		(Hirschfeld, Feldker, & Zwitserlood, 2012)
90		(Kovalenko, Chaumon, & Busch, 2012)
91		(Schendan & Ganis, 2012)

No.	Decade	Study
92		(Butler, Mattingley, Cunnington, & Suddendorf, 2013)
93		(Diéguez-Risco, Aguado, Albert, & Hinojosa, 2013)
94		(Giglio, Minati, & Boggio, 2013)
95		(Olivares, Saavedra, Trujillo-Barreto, & Iglesias, 2013)
96		(Proverbio, Azzari, & Adorni, 2013)
97		(Riby & Orme, 2013)
98		(Võ & Wolfe, 2013)
99		(Baetens, Van der Cruyssen, Vandekerckhove, & Van Overwalle, 2014)
100		(Balconi & Vitaloni, 2014)
101		(Boutonnet et al., 2014)
102		(Lensink, Verdonschot, & Schiller, 2014)
103		(Li & Lu, 2014)
104		(Manfredi, Adorni, & Proverbio, 2014)
105		(Mudrik, Shalgi, Lamy, & Deouell, 2014)
106		(Proverbio, Calbi, Manfredi, & Zani, 2014)
107		(Schleepen, Markus, & Jonkman, 2014)
108		(Dominguez-Martinez, Parise, Strandvall, & Reid, 2015)
109		(Dyck & Brodeur, 2015)
110		(Gao, Hermiller, Voss, & Guo, 2015)
111		(Kaczer, Timmer, Bavassi, & Schiller, 2015)
112		(Khushaba, Greenacre, Al-Timemy, & Al-Jumaily, 2015)
113		(Küper, Liesefeld, & Zimmer, 2015)
114		(Maffongelli et al., 2015)
115		(Ousterhout, 2015)
116		(Proverbio, Gabaro, Orlandi, & Zani, 2015)
117		(Schendan & Ganis, 2015)
118		(Zani et al., 2015)
119		(Zhou et al., 2015)
120		(Hoogeveen, Jolij, Ter Horst, & Lorist, 2016)
121		(Niu et al., 2016)
122		(Rojas et al., 2016)
123		(Yinan Wang & Zhang, 2016)
124		(Gui et al., 2017)
125		(Kiefer, Liegel, Zovko, & Wentura, 2017)
126		(Mandikal Vasuki, Sharma, Ibrahim, & Arciuli, 2017)
127		(Ortiz, Grima Murcia, & Fernandez, 2017)
128		(Pergola, Foroni, Mengotti, Argiris, & Rumiati, 2017)
129		(Savic, Savic, & Kovic, 2017)
130		(R. W. Y. Wang, Kuo, & Chuang, 2017)
131		(Bouten, Pantecouteau, & Debruille, 2018)
132		(Yi, Chen, Chang, Wang, & Wu, 2018)

Study design

Experiments and factors.

Out of 132 papers, 28 had reports on two or more experiments, and 19 of those had two or more experiments that fit criteria for this review. There were 17 papers with two experiments, and two papers with three experiments. Among them, in three papers (2.27% of the total sample) one of the experiments was replicated, with or without modification, in another experiment. One additional paper presented a replication and extension of a previously published study. The remaining papers presented discrete experiments, some of them with the same participants and/or including the same factors. This made classifying studies difficult, because equivalent designs were presented as two experiments in one paper, or one experiment in another paper. For example, a 2x2x2 repeated measures design could also be presented as a set of two 2x2 experiments with the same participants whose results were compared. For the purposes of our review, we described all studies the way the authors themselves had described their methods. Still, it is worthwhile to note that this decision affected the number of statistical analyses per experiment, which was smaller when results were presented as separate experiments. The only exception to our rule was one study in which data from two experiments was analyzed in a single joint main ANOVA, so they were treated as one experiment with two groups.[6]

About a third of papers (34.09%) had designs with one experimental factor, while the remaining two-thirds of the sample had two or three factors (excluding electrode factors). The tasks in which the N400 was measured were very diverse and some diverged from the traditional N400 paradigm. However, all papers explicitly interpreting a component as the N400 were included in this review.

Trial structure and timing.

Describing trial structure and timing is critical in ERP research (Donchin et al., 1977; Keil et al., 2014; Picton et al., 2000), since ERPs can be affected even by slight changes in these parameters. We investigated frequencies of using two common practices – jittering interstimulus/intertrial intervals and using techniques to prevent overt response from overlapping with the ERP epoch. These approaches are not applicable to all analyzed studies, but they are easy to incorporate in a typical N400 paradigm.

Regarding jittering, it was possible to determine if it was incorporated into trials for 125 papers (94.70% trials). The relevant information about timing was not provided in six papers (4.55%), and it was not possible to determine whether jittering took place due to unclear wording in one more study (0.76%). In addition, stimuli timing was variable due to other factors (e.g. self-paced trials) in 11 papers (8.33%), so jittering was not necessary.

[6] This study is not included in the 19 papers with two or more experiments previously described.

Among the remaining 114 papers, jittering was used in 30.70% cases. In one study, jittering was implemented in one of the two experiments, but not in the other.

About a half of all studies (52.27%) did not include any measures to prevent overlap between behavioral and ERP response. About a quarter (26.52%) of studies had designs in which there was no overt response to stimuli used in the N400 analyses, either because overt responding was not required or because the participants responded to other stimuli. Finally, a cue for participant to respond only after the ERP time window has passed was incorporated into the procedures in 21.21% of all papers, or 28.87% studies in which an overt response was required. There were no papers flagged as inconclusive, since it was considered that overt replies and cues to respond were used only if they were explicitly described in the section on trial structure.

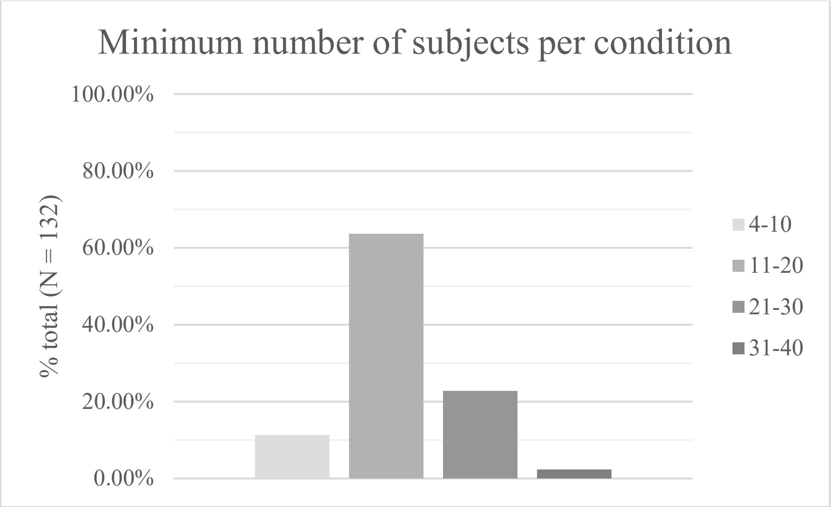
Sample size

According to publication guidelines, sample should be sufficiently large to allow enough statistical power to detect experimental effects of interest (Picton et al., 2000). This is not always easy to predict in ERP research, but a recent study has demonstrated that increasing sample size can dramatically improve statistical power in ERP research (Boudewyn et al., 2018).

In the papers we analyzed, sample size was mostly reported unambiguously – only one study was flagged as inconclusive, because its sample description and degrees of freedom in analyses did not match. However, there was one problem worth noting: in a small number of papers, sample descriptions did not allow determining whether the sample size was given with or without excluded participants. In these situations, we relied on degrees of freedom in statistical analyses to draw conclusions about sample size.

The number of participants recruited for studies varied between 5 and 66 (M = 20.44, SD = 8.35[7]). Most papers (60.61%) had an experiment with 11-20 subjects, while five studies (3.78%) had even fewer than 10 participants. The most frequent sample sizes were 20 (in 12.12% papers) and 16 (11.36% papers).

[7] When a paper had multiple experiments that differed on a numerical variable, their average was used for calculating the variable's mean and standard deviation.



Graph 2.1. Frequency chart of the smallest number of participants that was averaged in a study. Percentages are relative to the grand total of all papers (N = 132).

Although most studies we analyzed had within-subjects design, there were still experiments with two or more groups, so sample size per group is more telling than overall sample size. Since some papers had both within-subjects and between-subjects analyses, we examined the smallest number of participants that was averaged together for an analysis in each paper. The number of participants per group varied little less than the total sample sizes: there were between 4–40 participants (M = 17.48, SD = 6.25). The most common were studies with 11–20 participants per group (63.54%, see Graph 2.1). Sixteen studies (11.36%) went below 10 participants in their smallest group, while less than a quarter of all studies had more than 20 participants per condition: 22.37% papers had between 21–30 participants, and only three papers (2.27%) described experiments with more than 31 participants in all conditions.[8]

Number of trials

In addition to sample size, the number of trials averaged together is an important factor influencing statistical power. While there is no straightforward answer to the question of how many trials are needed for a visual image N400 study, it is known that increasing the number of trials can substantially increase power under some conditions, especially when it is in medium range (Boudewyn et al., 2018). Given the relevance of this parameter to signal-to-noise ratio, current publication guidelines (Keil et al., 2014) recommend specifying the number of trials presented in each condition, the average number and range of trials per

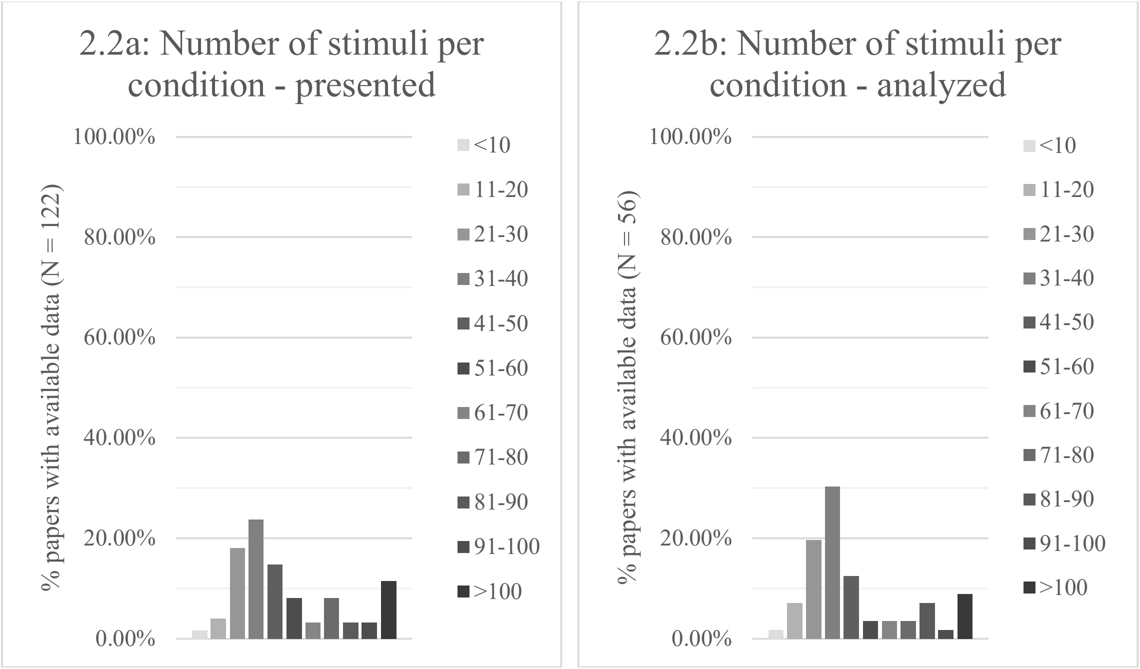
[8] In four of the studies, the lowest N per group came from an additional analysis, and not the main ones. Nevertheless, these studies had 4, 5, 10 and 23 participants, and neither the range (5-40 after excluding these papers), nor other parameters changed substantially when these papers were excluded.

condition left after rejecting trials, as well as whether the numbers differ substantially among experimental conditions or participant groups. Similar instructions were part of the previous version of the guidelines, as well (Picton et al., 2000).

Presented trials.

Study designs were, in general, described in a way that allowed calculating how many trials were presented in total, with only two studies (1.52%) labeled as inconclusive. One more study was not considered here because the trials count depended on participants' responses. As expected, among the remaining 129 papers, total trial count varied greatly depending on the study design, (range: 12–1584, M = 303.95, SD = 246.26). While the total trial count does not say much about a study on its own, it is evident that some studies had a very low trial count that would not provide enough trials per situation even for simple designs, while others were quite demanding for the participant.

Number of trials per situation is a more telling parameter for assessing data quality and statistical power. As with the number of participants per condition, we used the smallest number of trials in a condition when there were differences between experiments or within one experiment. The number of trials per condition was mostly described in an unambiguous way: two studies were marked as inconclusive (1.51%), although it was necessary to make calculations to deduce this information in some papers. In addition, 8 papers (6.06%) were not considered because trials were sorted into conditions based on participants' responses. Among the remaining 122 papers, the number of stimuli per situation varied considerably – between 6–400 (M = 60.78, SD = 51.57). Studies with 40 (16.39%) and 30 (12.30%) trials per condition were the most common, while all other individual counts were much less frequent. Majority of studies had between 20–50 trials per condition: 18.03% papers had 21-30 trials, 23.77% papers had 31–40 trials, and 14.75% papers had 41–50 trials per condition. Seven papers (5.74%) had fewer than 20 trials per condition, 26.45% papers were in 51–100 trials range, while 10.74% papers had even more than 100 trials per condition (Graph 2.2a).



Graph 2.2. Frequency chart of the minimum number of trials per condition. *Graph 2.2a:* The smallest number of trials presented per condition. Percentages show the proportion of the 122 papers in which this information was obtainable (i.e. the number of stimuli in a condition did not depend on participant response and it was presented in an unambiguous way). *Graph 2.2b:* The smallest of average numbers of trials left after all trial rejections. If information for individual conditions was not available, an overall average for all situations was used for approximation. Percentages are relative to the total count of studies for which this information was available (N = 56).

Analyzed trials.

While knowing how many trials were presented is important, it is probably even more important to know how many trials were averaged together in the analyses, since a portion of trials is usually rejected in ERP research, and its proportion can be appreciable under some circumstances.

Contrary to the publication guidelines, this information was frequently absent, and, when it was reported, this was done in many different formats, in most cases below the standards recommended in the publication guidelines, which made summarizing results difficult. More specifically, authors of 40.15% studies did not report how many trials were averaged together, and another 2.28% authors reported this information only for some experiments or analyses, and not for others. In addition, for 8.33% papers, it was not possible to extract this information even though it was reported: in 3.79% studies, rejection due to artifacts and overt response was presented separately making calculating total rejection rate impossible because of the potential overlap, and 4.55% of studies were marked as inconclusive due to ambiguous wording.

Apart from that, in 3.03% papers, the authors did not report discarding any trials due to artifacts or overt response. This leaves 61 (46.21%) papers in which trials were rejected and the number of remaining trials was reported. Authors of 13.64% of all papers reported the average number or percentage of rejections for each condition, along with the range of trial counts or the lower threshold for excluding a participant, which was the category closest to recommendations for good practice. The most common practice was to report only the overal average number of trials per condition (18.18% papers). Other authors reported average numbers for each condition (6.82%), only threshold for rejection or a minimum number of trials that was averaged together (5.3%), or overall average number of trials per condition along with the threshold or minimum (2.27%).

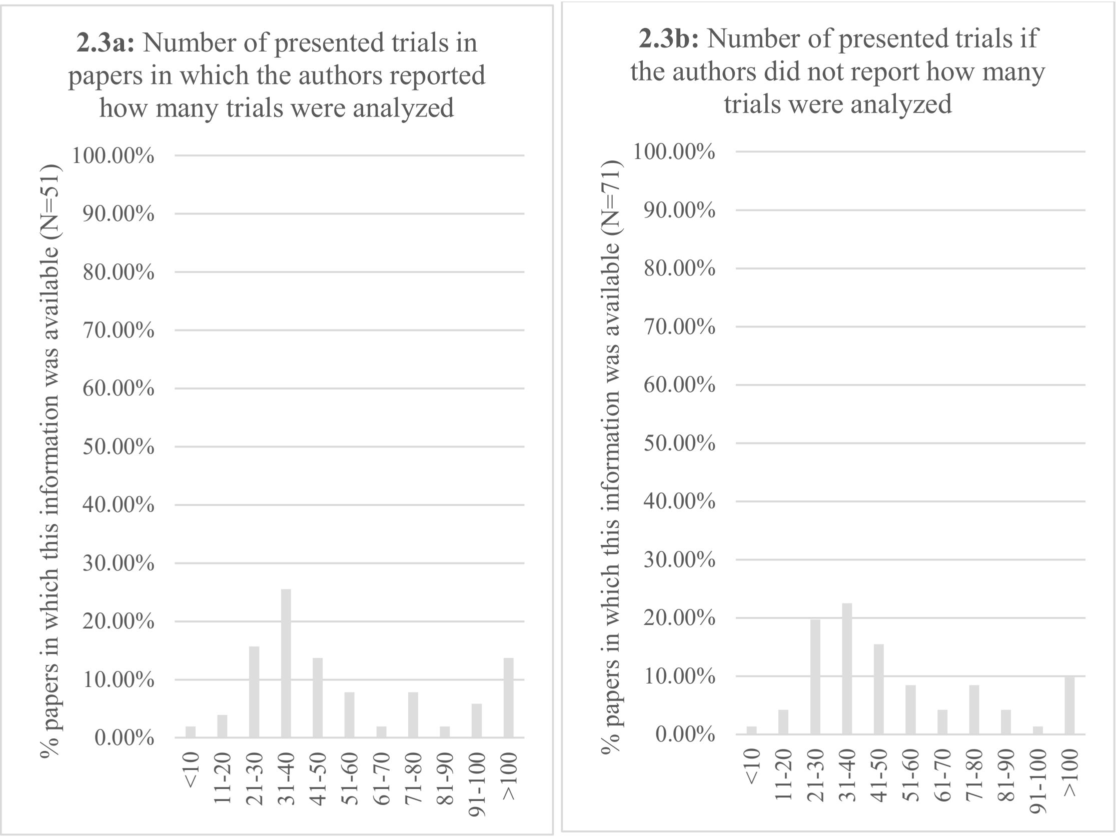
It was possible to extract the lowest average number of trials per condition, or, if not, to approximate it with the overall average for all conditions in case of 56 papers (42.42%). These papers showed that the smallest average number of trials that was averaged together varied between 4.60-182.95 (M = 50.32, SD = 36.33). The most common values were between 31-40 (30.36%), while 21-30 (19.64%) and 41-50 trials (12.5% papers) were also relatively frequent. Other options were less common. In total, 8.93% papers went below 20 trials per condition, whereas 28.57% papers had more than 50 trials on average in each condition, with 8.93% of papers which had more than 100 trials per condition (Graph 2.2b).[9]

These results could be biased – it is possible that the studies that had more trials in analyses reported this information more frequently. This is impossible to measure directly, but we tried to assess this possibility indirectly. To do that, we compared the number of trials that was presented to participants in two groups of studies - the ones that did and did not report the minimum number of trials averaged per condition. Studies in which the number of presented trials per condition could not be calculated were excluded from this comparison.

Distributions for both groups of papers are shown in Graph 2.3. The two groups were statistically compared using Mann-Whitney U test implemented in JASP 0.8.4.0. The results showed that stochastic difference between the two groups was not statistically significant ($Mdn_{reported}$=44, $n_{reported}$=51, $Mdn_{not\ reported}$=43, $n_{not\ reported}$=71, U=1742, p=0.721).

Our test is indirect and visual similarity or high p-value do not allow accepting a null hypothesis, but we can at least conclude that we did not find evidence of bias in reporting and that omission of the information on the number of stimuli from most papers could result from oversight.

[9] In addition to average number of trials that was analyzed, we were interested in what was the low end of the range that was tolerated. Because this was not reported in most papers, minimum number of trials per condition and threshold for eliminating participants were grouped together. In total, 29 studies (21.96%) reported either of these values, and they ranged between 5-90 (M = 30.75, SD = 19.93SD).



Graph 2.3. Distribution of the number of presented trials in two groups of papers: (3a) papers in which the authors reported how many trials were averaged together, and (3b) papers in which the authors did not report how many trials were left after trial rejection.

Equipment and software

Recording instrumentation and software used in ERP research vary greatly in properties which are crucial for evaluating, comparing or reproducing a study (Keil et al., 2014). Therefore, it is recommended that the recording equipment and analysis properties are described in detail.

For purposes of this review, we focused on one property of the equipment: were producers and models of instrumentation and software reported in the analyzed papers. While the guidelines do not specify that all equipment and software should be named (exceptions are amplifier and some types of software), this information can be useful to readers. It allows an informed reader to deduce some information when it was omitted, as well as to notice potential mistakes in reporting during review process. In addition, it can help researchers who are replicating a study or setting up a new laboratory with plans to build on an existing body of research.

Three quarters of papers (75.00%) provided at least some information on **hardware**.[10]

Amplifiers: In case of amplifiers, 40.15% papers did not provide any information on the producer or model, and in two additional cases, the information was not clear due to wording or conflicting details. In the remaining 78 papers, 17 different manufacturers were mentioned, but most of them only in a few papers. The most frequent producer was mentioned in 25.64% papers, in which it was used alone or in combination with other competitors for different experiments. Another four competitors appeared in 10-14% papers (48.72% in total), so the top 5 producers were used in 74.36% of studies.

Cap: About a half of all papers (46.21%) did not provide information on the type of cap (or an alternative electrode placement solution). In addition, in two cases, it could not be determined which cap was used due to wording or conflicting details. In the remaining 69 papers, there were 11 cap producers in total. The most frequent one was mentioned in 34.78% papers, and the next 4 competitors were referenced by 53.61% papers (10-17% each). Thus, top 5 producers appeared in 88.39% papers, while the remaining options appeared only in one or two papers.

Other hardware components: Most papers did not describe any other equipment specific to EEG/ERP recording (apart from information on electrode material, which is not covered in this report). In 10 papers (7.58%), information was provided on electrodes, ADC card, gel, electrode location digitizer, or a specialized monitor for presenting stimuli.

Details on **software** were less frequent than details on hardware components, with at least some information being provided in about half of all papers (49.24%).

Presenting software: Software for stimuli presentation was reported by 29.55% papers. In the 39 papers that provided this information, 12 different software packages were mentioned. There was one leading option, used by 43.59% authors, and the top 4 solutions were used in 79.48% studies. All other programs were referenced only in single papers.

Acquisition: Approximately one out of five papers (18.18%) reported information on software used for EEG recording. Granted, this information could likely be assumed based on amplifier details, but, as it has been described above, this information was also frequently omitted. These 24 papers included reports on 13 distinct acquisition software packages. Two

[10] This analysis does not make a distinction between cap and amplifier models by the same manufacturer. It would add further variability to results, and additional papers would be placed into the category of those that do not report these items. On the other hand, in some cases, it was possible to make further assumptions about hardware components that were not described based on the ones that were. This was the case with components that were made by the same manufacturer and that can only be used together. However, we did not include such inferred information in our report. A reader, or even a reviewer, does not necessarily have detailed knowledge about the equipment that was used in each study, and, in any case, there can be several distinct models by the same manufacturer that can be paired with the reported item.

programs were referenced by five papers each, and two more by two papers each, while the remaining options appeared only in single papers.

Processing: Programs for signal processing were the most frequently reported of all software. However, this information was provided in a third of papers (33.33%), and among these papers, there was one study flagged as inconclusive. Furthermore, among the remaining 43 papers, in 6 studies (13.95%) it was specified which software was used only for some processing steps, and not for others. In these 43 papers, 13 different software packages were referenced. The most frequently used one was mentioned in 30.24% papers, and top 3 options were used in 69.46% studies, alone or in combination with other programs.

Statistical analysis: Information on software used for statistical analysis was rarely provided. In 85.61% cases this information was not provided, and one additional study was flagged as inconclusive. The remaining 18 papers referenced 7 software packages, one of which was prevailing with references in 11 (61.11%) papers, while its competitors were reported in 1-3 papers each.

Other software: Three additional types of software were included in 13 of the analyzed papers. Software for advanced EEG data analyses (e.g. localization) was referenced in nine papers. Data visualization software was described in five papers (it is noteworthy that EEG topography maps were shown by about a half - 48.48% papers). Finally, in one paper, the authors reported using software for time-locking stimuli.

In summary, information on equipment models and software choices was frequently omitted from papers included in this review, including the types of information that are required by the publication guidelines. A quarter of papers (75.00%) do not contain any details on hardware makes or models, and about a half of all papers (50.76%) do not contain any information on software used. The most frequently reported item (and one required by the publication guidelines) was amplifier manufacturer. It was reported in about 60% of cases, and even then, information on the specific model was not always provided. Regarding other instrumentation, cap manufacturer was stated in about a half of papers (53.79%), and information on producer or model of other hardware components, such as electrodes or location digitizer, was available in a handful (7.57%) of papers. Regarding software, there was no agreement on which types of information should be provided: different papers reported different types of software, and reporting rates for software categories ranged between 14.39% for statistical analysis programs and 33.33% for signal processing packages. In addition to papers that do not report any information, in a few papers, issues in reporting didn’t allow determining what equipment or software was used.

Among papers from which information on hardware and software could be extracted, we looked at the amplifier and cap manufacturers (not specific models), as well as stimuli presenting, data acquisition, signal processing and statistical analysis packages (not versions). In all these cases, the same pattern could be observed. A variety of options was available for each item: there were more than 10, and up to 17, categories for all variables, except for statistical analysis software, for which 7 different options were mentioned by a total of 18 papers. Despite that, in all cases, there were only a few dominant options (between 1-5), chosen by a great majority of authors.

This shows that variability of both hardware and software bwas limited. Admittedly, it would be somewhat larger if individual models of caps and amplifiers had been separated.[11] However, the most frequently chosen options allow a fair amount of flexibility in recording and analysis decisions. Thus, variability that was found in some processing and statistical analysis choices cannot be attributed only to limitations of the instrumentation and software that was used.

Recording and processing

Impedance.

The appropriate way to handle electrode-scalp impedances depends on the recording apparatus, primarily on amplifier input impedance. Still, impedances can affect data quality and statistical power even in recordings with the more robust, high-input impedance amplifiers (Picton et al., 2000). This is particularly an issue with low-frequency components such as the N400, which are more sensitive to skin potentials – a problem that cannot always be remedied with high-pass filters (Kappenman & Luck, 2010; Picton & Hillyard, 1972). Therefore, publication guidelines (Keil et al., 2014; Picton et al., 2000) instruct that recordings with high impedances should be interpreted with caution and that suitable information on impedances should always be provided. Depending on the apparatus, it will be either electrode impedance higher margin, range of impedances or an alternative indicator of data quality (Keil et al., 2014).

In our sample, 7 out of 10 (72.73%) papers reported information on impedances, 24.24% papers did not report any measures of data quality, and 3.03% papers were flagged as inconclusive due to unclear wording. The papers that did report information on impedances, gave upper thresholds in almost all cases, although information on impedances range was given in one case, and two papers gave approximations ("usually less than 3 kΩ" and "below 15 kΩ at the beginning").

Among the 96 papers that reported impedances, 50% papers set the threshold of tolerance at 5 kΩ and one more study did so at 6 kΩ. The second most common margin was 10 kΩ, which was set in 9.09% studies. Various thresholds lower than 5 kΩ were implemented in 4.55% cases, while thresholds higher than 10 kΩ (15-50 kΩ) were reported in 6.82% papers. In one paper, two different recording systems were used for different experiments, so different thresholds were set – below 5 and below 50 kΩ.

Since adequacy of impedance threshold choice depended on amplifier input impedance, the papers were divided into two groups based on whether the amplifiers that were used were advertised as high-input impedance by their manufacturers. Papers that did not report amplifiers were excluded from this analysis.

[11] In addition, more papers would be placed in the category of those not reporting analyzed items.

In total, there were 50 papers with traditional, low-input impedance amplifiers. Understanding that impedances are an important factor in low-input impedance recordings was widespread – 42 (84%) papers included information on impedances, and in most cases the threshold was strict. In these 42 studies, most authors lowered impedances below 5 kΩ (73.81% papers) or even lower, below 2 or 3 kΩ (4.76% papers). However, 12% of authors did not report impedances thresholds, and in 4% of cases, the report was labeled as inconclusive on this parameter. About 20% of authors were more liberal when it comes to impedances: 9.52% tolerated impedances above 10 kΩ, up to 25 kΩ, while 10 kΩ was the threshold in 11.90% of studies.[12] When this information is coupled with modest average trial counts and sample sizes described earlier, these studies could probably benefit from more careful scalp preparation in terms of increase in statistical power.

There were 28 papers in the high-impedance group. Most of them (57.14%) did not have any information or impedances or alternative data quality parameters, which shows that there was not enough recognition that appropriate data quality indicators should be provided even with high impedance recordings. Among the remaining 12 papers, nine reported impedances below 10 kΩ: in three papers, the bar was set at 10 kΩ, five papers at 5 kΩ, and one paper reported a threshold even lower than 5 kΩ. Only three papers reported high impedances – in two papers only the higher margin of 50 kΩ was given, while in one paper, the range of impedances (10-50 kΩ) was reported.

Recording montage (active sites).

Despite instructions that recording sites should be described using standardized nomenclatures (Donchin et al., 1977; Keil et al., 2014; Picton et al., 2000), a large proportion of papers did not include this information. Namely, 53.79% papers contained details on all electrode sites, and one paper included information on all sites but one. In addition, 5.30% papers used nets that have fixed layouts (as opposed to caps, in which electrodes can be placed or taken off). Although these layouts were not clearly presented in these papers, information on them could be at least found online. These three groups together gave a total of 59.85% papers with (mostly) known information on recording montage.

On the other hand, 40.15% papers did not have clear reports. In 32.58% papers, the recording montage was not described. The description was labeled as inconclusive due to conflicting information in 6.82% papers, usually between graph, electrode list and electrode count. Finally, in one paper with several experiments, the electrode layouts were not described, but in one of the experiments, a geodesic net was used, so the layout could be found online.

When the recording montage was not shown, it was not only impossible to determine exact sites, but also overall active channel counts. Firstly, reference sites were treated in

[12] It is noteworthy that the latter is still acceptable according to publication guidelines, even if not enough for eliminating skin potentials. Ideally, that would require lowering impedances to 2-3 kΩ (Picton & Hillyard, 1972; Picton et al., 2000).

different ways in different papers. In some cases, mastoids were included in the total electrode count, while in others, this was not the case. Likewise, some, but not all, authors that used average reference counted the additional channel gained by turning the previous reference into an active site. This inconsistency was registered even in papers by the same author teams. Secondly, in some studies, not all slots in a cap were used – for example, the authors reported using a 32-channel cap and 28 or 29 sites. When this is applied to the papers in which the recording sites were not specified, we can conclude that it is not certain what the reported electrode counts represent. For purposes of this review, we reported the electrode counts as they were given in the original papers when sites were not listed, and we counted only active sites whenever possible.

There were 34 different electrode counts in 123 papers in which the number of recording channels was not inconclusive. They varied between 1–144 (M = 46.33, SD = 36.08). Small montages, with up to 10 electrodes, were not frequent (11.38%), and only one active site was used for recording in one study. Likewise, very large montages, with more than 64 sites, were used in 13.82% papers. In about half of all studies, recordings were made using either 32-channel or 64-channel caps in which most locations were utilized. Namely, there were between 28–32 active electrodes in 25.20% studies, and between 60–64 active sites in 26,01% papers. Common options were 64 sites (19.51% papers), 128/129[13] (11.38%), 29 (9.76%), 32 (6.50%), and 20 sites (5.69%). Other montage sizes were represented in fewer than 5% of papers.

Montages with the same electrode counts can be further broken down by sites that were used. Among 79 papers that provided information on which sites had been used, most electrode counts could be divided into more than one, and up to four, different layouts. In total, there were 50 different recording layouts in these 79 papers, 37 of which were used in only one study. Five montages were used in more than three studies, and the most frequently used montage appeared in six papers (7.59% papers that reported this information and 4.55% of the total sample).

The most common recording montages were:

(1) a geodesic net of 129 channels with fixed layout; it was used in 6 studies; an example can be found in the paper by Dominguez-Martinez, Parise, Strandvall, & Reid (2015); the same authors appear in 2/6 papers;
(2) a montage of 29 electrodes: it was used in 5 studies; an example can be found in the paper by Eddy, Schmid, & Holcomb (2006); the same authors appear in 4/5 papers;
(3) a 64-channel cap with all sites used; it was used in 5 studies; an example can be found in the paper by Bramão et al. (2012); the same authors appear in 2/5 papers;

[13] Some geodesic nets with 128 sites by the same manufacturer were listed as having 128 channels, while others were listed as having 129 channels, depending on the reference point and whether it was included in the electrode count.

(4) a montage of 20 electrodes: it was used in 4 studies; an example can be found in the paper by Olivares, Iglesias, & Bobes (1999); the same authors appear in all four papers;

(5) a geodesic montage with 26 electrodes: it was used in 4 studies; an example can be found in the paper by Wicha, Moreno, & Kutas (2003); the same authors appear in all four papers.

As it can be seen from the description above, the same layouts came mainly from the same laboratory.[14] Two of these five montages were always used by the same author teams, and one montage was used by the same authors in 4/5 cases. The remaining two montages were used by the same authors in two cases (out of five and six), but one of them was a geodesic net with an inflexible layout, and one was a cap in which all sites were used.

Part of the observed variability can be attributed to using different cap models, but part of it can be ascribed to allocating electrodes differently within the same cap when not all sites are used. This variability can make analyses which encompass large numbers of electrodes less comparable, but it doesn't necessarily affect traditional analyses of a smaller set of electrodes, which can be applied to some commonly recorded sites.

Reference and re-referencing.

Choosing a reference is an important issue, without an ideal solution (Luck, 2005, 2014). The guidelines (Keil et al., 2014; Picton et al., 2000) recommend that online reference should be given, as well as the digital one if re-referencing is used. Keil et al. (Keil et al., 2014) additionally stress that in case of average reference, this includes showing all sites that were included in the average. The authors are encouraged to include graphs with the reference typical for their field if they are using a different reference (Picton et al., 2000). Physically linking electrodes is not recommended (Keil et al., 2014; Picton et al., 2000) because it leads to distortion of voltage distribution over scalp (Miller et al., 1991).

In this review, we were interested in the reference that was used in N400 analyses, be it online or digitally computed. When it comes to reporting clarity, only three papers (2.27%) were labeled as inconclusive, and three more didn't include reference information. However, in 8 papers (6.06%), a mastoid or linked reference was used, but it was not specified, or not clear, which of several options was implemented. Furthermore, in 8.33% of all studies, the average reference was used, but it was not reported which sites were included in this reference (more details below). Including these two issues, it was not possible to reproduce exact reference based on 18.93% reports.

There were several other minor issues in reporting, but one of them stood out. The term "linked mastoid/earlobe reference" was used to refer to both physically linking

[14] Some authors were represented with multiple publications in our sample of studies. Two authors were authors of more than 5% of all papers (6.06% and 7.58%), and two more had 4.55% publications in the sample. Their choices influenced results more when the variability was high, and they were consistent in their choices. When this is the case, it is noted throughout the paper.

electrodes and the average of physically separate sites, sometimes without specifying which of the two had been implemented. For purposes of this review, if re-referencing was described, we assumed that a linked reference was obtained offline by averaging. Otherwise, if recording with a linked reference was mentioned, it was assumed that the authors referred to physical linking.

In addition to issues in reporting, in a handful of papers, re-referencing from a linked recording was described – this would not lead to the same result as re-referencing from a single reference due to the above-mentioned effects on voltage distribution.

Which references were typically used in the analyses of 126 studies which reference information was given? In five studies (3.97%), the authors used multiple different reference points to compare results of N400 analyses. In all these papers, mean mastoids were combined with other references. It was usually the average reference (4/5 papers), but two papers also included Cz reference, which not an optimal choice given central distribution of the N400 (Luck, 2014).

Regarding individual options, most papers included some variation of mastoid or earlobe reference (74.60%, including papers with multiple references). Mastoid references were used in 56.35% of these studies. The most common among mastoid references was mean mastoids (34.92%). Other options were left mastoid (6.35%), right mastoid (0.79%), linked mastoids (7.14%) and sum of mastoids (2.38%). Finally, in six papers (4.76%), it was not specified or not clear which of these options was chosen. Most earlobe references were linked (15.87%), and the remaining options were right earlobe (0.79%) and unspecified/unclear earlobe reference (1.59%), giving a total of 18.25% earlobe references among papers that were not labeled as inconclusive.

The second most common group of references was the average reference. It was employed in 27 papers (18.25%), including four papers in which a mastoid reference was also used. This reference strategy comprised a variety of distinct references, as data in these 27 studies had been recorded with at least 14 different montages, varying in size between 19 and 144 sites. This variability presents an obstacle to comparing the results of studies using the average reference (Luck, 2005, 2014). The exact number of distinct references could not be established because in almost half of these papers (44.44%) the montage on which an average reference was based could not be extracted. Among the 15 papers (55.56%) from which it was possible to extract the reference, 7 papers did not show recording layouts, but they did report using nets with fixed layouts that could be found online. In total, there were only 8 papers (29.63% average references) in which it was shown which sites were used to generate the average reference. On top of the issues of describing recording montages and their variability, electrode sets used to generate the average reference should be large and cover a large area of the head (Junghöfer et al., 1999), which was not the case in all of the evaluated studies. Given all these issues, it can be concluded that the average reference was not always applied nor presented in line with recommendations.

Besides mastoid/earlobe and average references, there were three other options, although rarely used. Tip of the nose was reference of choice in 5.56% papers in which reference was given, and balanced non-cephalic reference was used for analyses in 1.59%

papers. As mentioned earlier, Cz was used as a reference in 1.59% papers as well, but as an additional reference when mean mastoids were used.

All in all, while there may be advantages of other solutions in some cases, mean mastoids reference, or its nearby alternatives (mean earlobes, P9&P10), would be the best choice for visual N400 studies in most cases, given that it would allow comparing the results to majority of literature, and that it is preferable to the single or linked reference (Luck, 2014). The way the average reference is used in the existing literature, its costs outweigh the benefits. If a researcher would want to use it, they should pay more attention to the recommendations to consider its potential bias (Junghöfer et al., 1999; Keil et al., 2014; Picton et al., 2000), to include plots based on mastoid reference to enhance comparability with previous research (Picton et al., 2000), and to describe the montage on which the reference is based (Keil et al., 2014; Picton et al., 2000).

Filtering.

An adequate report on filters should contain information on the filter type, cut-off frequency with information on what kind of cut-off point is reported, and slope (Keil et al., 2014). Similar instructions were laid out in the earlier versions of the guidelines (Donchin et al., 1977; Picton et al., 2000), although Picton et al. specifying roll-off is only if the cut-off is close to the ERP frequency range. The problem of adequate reporting on filtering has been repeatedly discussed in other sources as well (Cook & Miller, 1992; Luck, 2005, 2014). Regarding specific choices, it is recommended that the final band pass should have half-amplitude cut-offs between 0.05-0.2 Hz (high-pass) and above 20 Hz (low-pass) (Luck, 2014) for this type of study, and 0.1 Hz was shown to be the optimal choice for N400 recordings (Tanner et al., 2015). Regarding roll-off, a relatively gentle slope (12-24 dB/octave) is recommended, especially for the high-pass filter (Luck, 2014). High-pass filters can be especially problematic, since filters with a half-amplitude cut-off of 0.3 Hz and higher decrease the N400 effect and instead introduce artifactual effects of positive polarity in P200 and P600 ranges (Tanner et al., 2015). Using notch filters is not recommended (Picton et al., 2000). In addition to these recommendations, we looked whether it was reported that high-pass filtering was applied on epoched data, which is not recommended because of edge artifacts that are produced this way (Keil et al., 2014; Luck, 2005, 2014), and whether digital filters were identical to offline filters, both of which could indicate misunderstanding of the effects of filters on signal.

High-pass and low-pass filters: cut-off.

Some filter cut-off information was provided by 97.73% of papers, and about 80% of papers reported both online and offline (if applicable) filter cut-offs.

Although cut-off frequencies were mostly reported, in most cases (71.21% studies) it was not indicated which point they represented – half-amplitude or half-power. Furthermore, among the 38 remaining papers, only partial information (for some of the filters) was provided in 11 (28.95%) studies. Most of the reported values (92.11%) referred to *online filters*. The type of online cut-off point was specified for about a third of online filters (high-pass filters: 33.65%, low-pass: 33.01%). Cut-off frequencies were given in 30 papers, and they were divided almost equally: half-power frequency was given in 16 papers, while half-amplitude point was reported in 14 papers. Time constant for the high-pass filter was given in the remaining five papers (without defining low-pass filter cut-off type), as well as in two papers in which low-pass filter cut-off type was defined. *Offline filter cut-off* type was reported in only three papers. In two of them it was a half-amplitude measure, and in one half-power.

Online high-pass filter cut-off point was omitted from 21.21% papers. Among the 104 papers that remain, settings were rather diverse. DC recordings were used in 8.65% studies, and cut-offs (both half-power and half-amplitude) for the recordings in which low-pass filtering was used varied between 0.001-1.05 Hz. There were 14 distinct filter cut-off points, the most frequent being 0.01 (27.88% papers that report online low-pass cut-off), followed by 0.05 (22.12%), 0.1 (11.54%), and 0.016 (9.62%). Since cut-off type and roll-off were not indicated in most studies (see below), it is impossible to tell more from these values. However, it can be noted that most papers (86.54%) reported values lower than 0.3 Hz, meaning that most amplifiers allowed recording N400 without filtering artifacts (Tanner et al., 2015)[15]. In one paper, two different cut-offs were used for different experiments. In this case, different equipment was used, and digital filtering was not used to make filters comparable. Tables with cut-off frequencies of all filter settings can be seen in Appendix E.

Online low-pass filter cut-off value was absent from 19.70% papers. In the remaining 106 papers, 18 different cut-off points between 20-256 Hz were mentioned. Since the half-amplitude point of a low-pass filter is higher than its half-power cut-off, all of these values fall within range acceptable for this type of ERP studies (≥20 Hz half-amplitude) (Luck, 2014). The most common settings included 100 Hz (31.13% of reported low-pass filters), 40 Hz (17.92%), 30 Hz (18.87%), and 70 Hz (9.43%) (either half-amplitude or half-power).

Offline high-pass filters were applied in 37 (28.03%) studies. Notably, when an offline high-pass filter was reported, in 59.46% cases, information on the online filter was missing. This was not the case when digital filtering was not mentioned. Like problematic filtering choices, this could, too, indicate possible misunderstanding of properties and effects of filtering. Nine different cut-off options were mentioned, between 0.01-1 Hz. The most common choice was the optimal choice (Tanner et al., 2015) – 0.1 Hz (16 papers, 43.24%). Other common options were 0.01 Hz and 0.3 Hz (five papers, 13.51% each), while 0.5 Hz and 1 Hz were applied in three studies (8.11%) each. Like above, an appropriate summary of this information would require knowing the cut-off type, and half-power values would be translated into somewhat lower half-amplitude values, but in 29.73% papers that included offline filters, the chosen cut-off values were potentially problematic (≥0.3 Hz). When online and offline filters are taken together, in 14 studies (10.60% of all papers), either online or

[15] In the study by Tanner et al., half-amplitude cut-offs were given, and some of the papers in our sample reported half-power cut-offs. Nevertheless, in case of high-pass filters, half-power cut-off is higher than half-amplitude point, so all these papers, and possibly a few others, have half-amplitude cut-offs below 0.3.

offline high-pass cut-off point was (at least potentially) outside of the recommended range. In addition, it was reported that digital high-pass filtering was applied on epoched data in one paper, and in two papers, no offline high-pass filtering was mentioned even though DC recordings were used.

Finally, *offline low-pass filters* were more common than offline high-pass filters – they were used in 57 (43.18%) studies. Like in case of high-pass filters, in 40.35% of cases when a digital low-pass filter was used, details on online filters were not given. There were 14 different cut-off points. The lowest cut-off point overall was 5 Hz, but it had been used for a peak latency analysis in one paper. Among the remaining papers, cut-off frequency still varied substantially, ranging between 5.5-100 Hz. Not unexpectedly, 30 Hz was the most common choice (49.48%). Cut-off at 20 Hz (14.03%) was the only other option implemented in more than 10% of studies, and all other options were less frequent. Like in case of other filters, it is impossible to tell precisely without other filter parameters, but in 19.30% studies, offline low-pass filters were potentially outside of the recommended range (<20 Hz half-amplitude). These papers make up 7.58% of all studies, which is also the percentage of all studies with potentially too narrow low-pass filters, given that all online filters were within the recommended band.

In addition to above-described issues, in two papers, filters with the same online and digital cut-off frequencies were applied, and in one paper, digital filter settings were more inclusive than online settings.

High-pass and low-pass filters: roll-off.

Unlike cut-off frequency, roll-off was rarely described. In case of *high-pass filters*, 81.82% papers reported neither online nor offline slopes, and one more study was labeled as inconclusive. In two studies, DC recording was used and there were no digital filters, so no high-pass filtering was reported. Furthermore, 19 of the 21 remaining studies included only partial information (either for online or for offline filters), and the remaining two studies only had online high-pass filters. Slopes were reported mostly for *offline* filters (17 papers, 80.95% roll-off reports). About a half of all offline high-pass filter reports (49.95%) included slope information. Two slopes were used in all these studies: 12 dB/octave (64.71% papers) and 24 dB/octave (35.29%). On the other hand, *online* slope was specified in only four papers (3.23% of all papers that did not report DC recordings), and it was either 3 dB/octave (two papers), 6 or 12 dB/octave (one paper each).

It was similar with *low-pass filters*: 79.55% reports specified neither online nor digital filter slope, and one additional study was labeled as inconclusive. Among the 26 remaining papers, only one specified both online and digital filter slope, and one more had only online low-pass filters. Out of these 26 papers, 19 (73.38%) included information on *offline* filters. This was a third (33.33%) of all papers in which digital low-pass filters were used. In three of these papers, offline filter slope was calculated based on Butterworth filter order. Digital low-pass filters had slopes of 24 dB/octave (eight papers, 42.11%), 12 dB/octave (seven papers, 36.84%) or 48 dB/octave (three papers, 15.79%). Online low-pass filter slope was reported in eight papers (6.06%) of the total sample, and it was relatively mild in all cases: 12 dB/octave (3 papers), 24 dB/octave (two papers) and 6 dB/octave (one paper).

Other filters.

In addition to high-pass and low-pass filters, notch filters were sometimes applied, although not frequently. Notch filters were used in 14 studies (10.60%) to eliminate electrical noise: a 50-Hz notch filter was used in 7.58% studies, and a 60-Hz notch filter was implemented in additional 3.03% studies. Aside from band-pass and band-stop filters, additional linear detrending algorithms were mentioned in four papers (3.03%). Only a handful of papers reported additional smoothing of ERPs for plotting.

To summarize, the percentage of papers which contained at least some information on filtering showed that its importance was widely accepted – only three papers (2.27%) did not mention filters. However, a more detailed analysis demonstrated that it was not as widely recognized what a sufficiently detailed description of filters should contain. Only about 80% of papers had online cut-off frequencies, and this information was frequently excluded when digital filtering was applied. Furthermore, most (71.21%) papers did not specify whether cut-off values described half-amplitude, half-power, or some other value. Both in case of high-pass and low-pass filters, only about 20% of papers specified any roll-off information, and even then, it was only partial information (not provided for all filters) in most cases. Roll-off was more commonly reported for offline filters (a third for low-pass, half for high-pass) than for online filters (<10% papers). All this information, in addition to the type of filter (not investigated in this report) is necessary to evaluate, compare and replicate filtering procedures. Thus, it is not possible to fully understand filtering procedures based on information provided in most papers analyzed in this review.

Among papers in which appropriate information was provided, most choices were in line with recommendations. There were some exceptions. In about 10% of the studies, either online or offline high-pass cut-off frequency was 0.3 Hz or higher, and in about 8% of studies, offline low-pass frequency was below 20 Hz (with the caveat that it was not specified what cut-off values represented). Notch filters were used in 1 out of 10 studies. Other issues were noted only in a few papers: using the same filter twice, applying a digital filter wider than the online one, applying high-pass filters to averaged data.

Baseline.

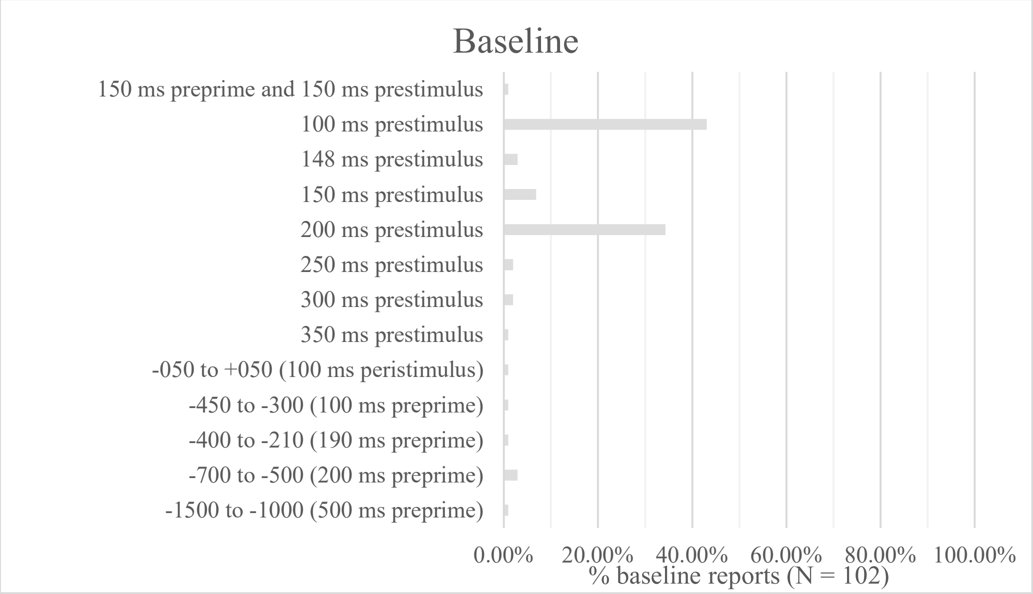
In this review, we focused on baseline used for amplitude measurement – other analyses, especially source analyses, may have different baseline requirements. According to guidelines, baseline period for amplitude measurement should be sufficiently long, at least 100 ms (Picton et al., 2000), method and data segments used for any baseline removal procedures should be specified in text (Donchin et al., 1977; Keil et al., 2014; Picton et al., 2000), baseline choice should be justified, it should not contain condition-related differences, and the entire baseline length must be shown in graphs (Keil et al., 2014). In the studies reviewed here, a prestimulus baseline was used in most cases. It is recommended that prestimulus baseline length should be at least 20% of the length of the poststimulus period, but not much longer than necessary, and that a length that is a multiple of 100 ms is preferable, unless increasing the baseline interval leads to more artifact rejections or including confounding ERP activity from the previous trial (Luck, 2014). Consequently, 100 ms or 200 ms is an optimal baseline duration for many N400 studies, with 200 ms being preferable if the measurement window extends beyond 500 ms or if other, later components are included. As it can be seen in sections on the N400 window and other components, both cases were common.

In 21.97% papers, baseline length was not presented, and one additional paper was marked as inconclusive, leaving 102 papers with baseline reports. Granted, in some cases in which baseline information was not provided, prestimulus epoch length was given. However, while these two values are usually the same, this is not necessarily the case, and they were not equal in some of the papers included in this review. Notably, in some data processing systems, baseline correction takes place automatically during epoching, so the researchers may not be aware that these two values can be different and that they should be presented separately. Additionally, some of the papers in which baseline length was given also didn't have entirely unambiguous wording, but it was could assumed with reasonable certainty that the epoch length and baseline length were equal.

These were not the only issues with baseline correction. Several other problems were registered, but their quantifying exceeds the scope of this review. However, they are listed here because they compromise ERP analysis or interpreting graphical representations of waveforms, and some of them were not scarce in the reviewed literature. In some papers, the graphs did not show baseline-corrected data, and this issue was registered both in papers in which baseline correction was reported and in papers in which baseline correction was not mentioned. This might be an issue in plotting, and not in data processing, but in either case, it would pose a problem. In the first case, the graphs can be misleading as the differences between lines are not accurately presented, and in the second case, the analyses could give wrong results if the difference between waveforms would be changed by baseline correction. In some papers, there was a considerable amount of noise in baselines or there were confounds causing waveforms to dissociate starting from the baseline period. In a few papers, the graphs did not show the entire baseline period, including cases where the baseline period was not show at all. In one case, the baseline period was reported to be longer than the prestimulus epoch it was based on, and in another paper, it was reported that baseline correction was not conducted, but a later description of measurement included baseline.

Out of the 102 papers in which baseline was described, 94 (92.16%) had prestimulus baselines, 6 papers (5.88%) had preprime baselines, one paper had a peristimulus baseline (-50–50 ms), and one paper had different types of baseline in different experiments. Among prestimulus baselines, 100 ms was the most common length (44 papers, 46.32%), followed by 200 ms (36.84%), and 148/150 ms (11.58%). Other options were 250, 300 and 350 (5.26% in total). These results are shown in Graph 2.4.

In conclusion, these results suggest that baseline is one of the aspects that require more attention when reporting on N400 studies (and likely other ERP studies, too). The specific points that could benefit from more scrutiny by authors and reviewers were baseline descriptions in text, representation of the baseline period and baseline correction in graphs, and presence of noise and confounds in the baseline period. In addition, some studies may benefit from extending the baseline from 100 ms, which was the most common option, to 200 ms. This could enhance amplitude measurement stability, especially if the N400 latency range extends beyond 500 ms (Luck, 2014).



Graph 2.4. Frequency of choosing different baseline lengths and positions. Percentages are relative to the total count of papers in which this information was available (N = 102). Two different baselines come from a study with two experiments which differed in the way baselines were calculated.

Poststimulus epoch.

Epoch length.

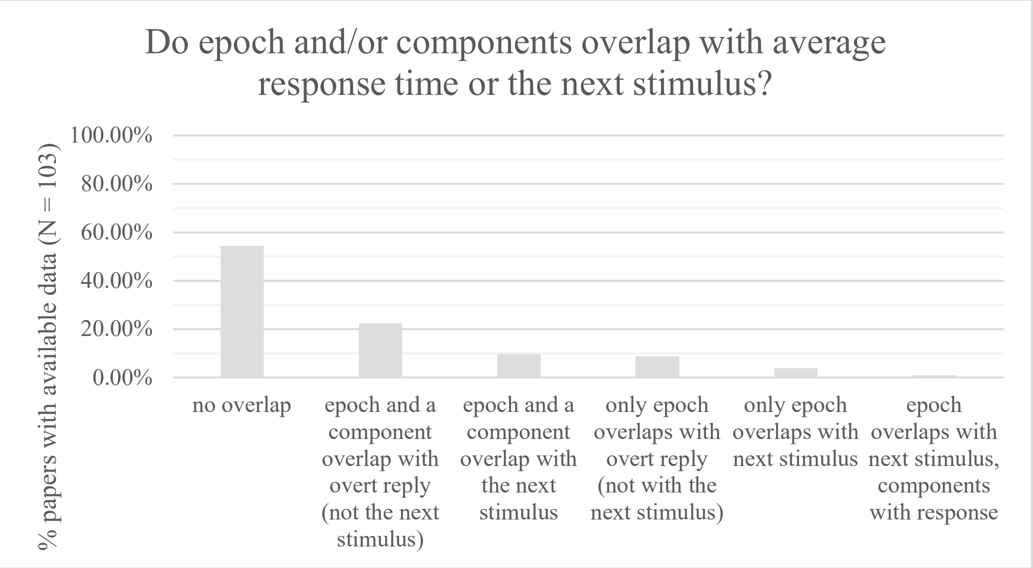
Poststimulus epoch length could be extracted from 83.34% papers – 12.88% papers didn't have this detail, and in 3.78% cases it was described in a way that it was not clear whether it included prestimulus period or not.

In these papers, 32 different epochs could be found. They were 550–2500 ms long, and the most typical lengths were 1000 ms (21.82% cases), 800 (11.82%) and 1200 (10.00%). These epochs were in most cases paired with a baseline shorter than the recommended 20% of the poststimulus period (Luck, 2014). Out of 82 papers in which both durations were given, baseline length was at least 20% of the poststimulus period in 22 papers (26.83%). Late epoch portions outside of the N400 window latency don't affect our component of interest, but many papers also included other, later components. When it comes to the N400 per se, post-N400 epoch length was of interest only to the extent that it could lead to more trial rejections based on artifact presence.

Appendix F includes frequencies of choosing individual epoch lengths and how they were matched with baseline durations.

Overlap with overt response or the next stimulus.

In addition to including occasional artifacts such as eye blinks, the epoch could also encompass ERP activity resulting from overt response to stimuli or the beginning of the next stimulus. This ERP activity does not necessarily pose a problem, but it is a potential confounding factor when it overlaps with time windows of interest. Because reaction time (RT) range and standard deviation were not always given, we inspected overlap between the average RT and epoch.



Graph 2.5. Distribution of studies based on whether epoch or analysis time windows included either average behavioral reaction time, or the beginning of the next stimulus. Percentages are relative to the total number of papers from which this information could be extracted (N = 103).

In 21.97% of papers, it was not possible to tell whether poststimulus epoch overlapped between the average RT or the beginning of the next trial because some relevant timing information (epoch length, RT or trial structure) was missing. Granted, in 41.38% of these papers, it was possible to at least determine whether there was an overlap with component time windows.

Among the remaining 103 papers, poststimulus epoch overlapped neither with the average RT nor with the next stimulus in 54.37% cases (see Graph 2.5). In 12.62% papers, there was some overlap, but it affected only the epoch, and not any analyzed time windows: in 8.74% papers, the epoch overlapped with the average RT, and in 3.88% it included the beginning of the next stimulus. Finally, in about a third of these papers (33.01%), there was overlap not only with the epoch, but also with some of the analyzed time windows: epoch and an analyzed component overlapped with the average RT, but not with the next stimulus in 22.33% studies, both overlapped with the next stimulus in 9.71% studies, and in 0.97% studies, the average RT fell within an analyzed time window, while the beginning of the next trial overlap with the poststimulus epoch.

Eliminating artifacts.

Even the earliest version of the guidelines (Donchin et al., 1977) specified that contamination of ERP recordings with artifacts must be addressed in studies, and the way it was addressed should be presented in reports. More recent guidelines give more specific instructions on how to describe artifact rejection (Picton et al., 2000) and correction (Keil et al., 2014), as well as on some aspects that should be considered when choosing the appropriate approach.

Accordingly, most papers (129, 96.97%) did mention implementing procedures to eliminate artifacts, and it was generally clear (125 papers, 94.70%) at least whether correction, rejection or some combination of the two was employed. Four papers (3.03%) did not mention artifacts, and another three papers (2.27%) were labeled as inconclusive regarding correction vs. rejection choice.

However, like in case of baseline correction, when stricter criterions were taken, more papers were shown to have issues. In 16.67% papers, it was reported whether correction or rejection was used, but no details were provided on the method of artifact rejection. Eight papers (6.06%) had procedure descriptions that were inconclusive, giving a total of 8.33% inconclusive reports together with the papers in which the correction vs. rejection choice was categorized as inconclusive. Consequently, a total of 95 papers (71.97% of the whole sample) had artifact elimination descriptions that allowed concluding anything more than whether correction, rejection, or some combination of the two was implemented. Moreover, additional 5.30% papers had details only for some of the artifact elimination steps, and some of the remaining papers provided insufficient details about the procedure, the most extreme examples being papers in which it was only stated that correction was regression-based or rejection threshold-based.

When it comes to reporting on artifact correction, 48 papers reported implementing it (36.36%), either alone or combined with rejection. Thirty-five (72.29%) of them included information on which algorithm was used to compensate for artifacts. Not all these studies provided more details beyond naming the type of correction. Eight papers (16.67%) did not report which approach was used, four (8.33%) only stated that correction was based on regression, and one study (2.08%) was marked as inconclusive.

Artifact rejection was used in some form in most papers (120 studies, 90.15%), but its descriptions were often incomplete. Namely, 22.69% of them had no details. In 9.24% cases, at least some steps were labeled as inconclusive. Some studies included steps that were either not described (3.33%) or they were vaguely described (7.50%), e.g. it was only said that rejection was based on a threshold. As a result, a total of 69 (57.50%) papers in which rejection was used, included descriptions of all rejection steps.

What approach was used in studies from which this information could be extracted? Regarding the most basic distinction – correction vs. rejection, rejection was a dominant choice. It was used by 62.40% of 125 papers that could be categorized by this criterion. It was followed by approaches combining rejection with correction (28.80%). Relying only on correction was rare (5.60%). In two studies (1.60%) different approaches were taken for different experiments. Contrary to the guidelines, authors of one study (0.80%) did not apply any measures to eliminate artifacts, and in one study, instead of rejecting the entire trial with an artifact, only the electrodes which exceeded a threshold were eliminated.

General approach to detecting artifacts could be either visual, based on a numerical threshold, on some more advanced algorithms, or a combination of these procedures. It could be identified for all steps in 88 papers, and for at least some steps in 95 papers. Out of these 95 papers, 35.79% studies relied on numerical thresholds, 12.63% relied on visual inspection, 10.53% used more complex algorithms, and 37.89% incorporated some combination of these steps (usually a numerical threshold for rejection and a correction algorithm). As noted in the previous section, in two studies, there were multiple experiments with different processing strategies, and in one study, no steps were taken to eliminate artifacts.

A list of specific strategies used in all studies would be long – there were 67 unique artifact elimination pipelines in the 95 papers in which at least some details were provided. Given that the description was only partial in some papers, this number could be slightly higher or lower depending on the missing details, but the difference could not be large as most papers with partial information diverged in the details that were provided. The pipelines varied from a simple base-to-peak threshold for epochs to elaborate procedures with several stages of rejection and correction.

We will briefly present which rejection and correction steps were common in these pipelines. There were nine different methods in the 35 papers in which artifact correction approach was named. The most common of them were independent component analysis - ICA (Jung et al., 2000) (31.43% artifact corrections), and two methods based on regression – developed by Semlitsch, Anderer, Schuster, & Presslich (1986) (22.86%) and by Gratton, Coles, & Donchin (1983) (20.00%). Other approaches were less frequent (up to 8.57%). On the other hand, there was a variety of rejection criteria in 69 papers in which all rejection steps were described. In 51 of these papers (73.91%), rejection criteria included some type of numerical threshold. The most common type of tolerance limit was maximum amplitude (base-to-peak). It was included in two thirds of the 51 rejection strategies relying on numerical criteria. The most common limits were 75 µV and 100 µV, although the value ranged between 50-200 µV. Peak-to-peak amplitude thresholds were also frequent (37.25% threshold-based strategies), and they ranged between 40-400 µV, within epoch or within a short window. The most common limit was 50 µV, although this reflects the preference of one laboratory which was represented in the sample with multiple publications and had a consistent approach to artifact rejection in most studies. Cut-off values based on variance/SD, step between two consecutive sampling points, and periods of low activity were also used. In most papers, thresholds were the same for all participants, but in 4 papers (7.4% threshold-based strategies), they were adjusted for each participant. Visual inspection was a basis for detecting all artifacts in 17 papers (26.64% rejection strategies), and it was one of rejection criteria in 4 papers (5.80%). In two papers, rejection was based on a more complex algorithm, ADJUST (Mognon, Jovicich, Bruzzone, & Buiatti, 2011) and FASTER (Nolan, Whelan, & Reilly, 2010). Other recordings (gyroscope, eye tracking) were used to detect EEG artifacts in another two papers, one of which was not in this group of 69 papers in which all rejection criteria were specified. Lists of all rejection and correction approaches can be found in Appendix G.

In addition to criterions for handling artifacts, in one paper, numerical criteria were set to identify and replace participants with relatively low signal-to-noise ratio (Pergola et al., 2017).

Order of operations.

Not all data pre-processing and measurement operations are linear, so the outcome may vary depending on the order of operations, and for some steps, one sequence is preferable to another (Luck, 2005, 2014). Therefore, in addition to describing individual data processing steps, it is important that the temporal order of these steps is presented in a clear way (Keil et al., 2014).

The evaluated papers had descriptions from which the order of operations can be extracted with a varying degree of certainty: from explicit accounts, through papers in which it was possible to presume that the order of steps was the same as the order in which they were described, to papers in which the processing steps were presented in such way that it was not possible to make any assumptions about their order.

In total, more than a half (53.79%) of the papers fell into this last group, meaning that it was not possible even to make an assumption about the order in which processing operations were conducted. Three common issues can be noted: (1) in some papers, the new reference after re-referencing was specified in the recording section together with online reference, and re-referencing was not mentioned later, in the processing section[16]; (2) a processing step that had likely taken place was not mentioned in the paper, so a reader couldn't be sure if it had taken place and at which moment (e.g. no mention of baseline correction or artifact handling); (3) the last step, averaging, was described first, in a sentence in which several other steps were mentioned as side points, in a way that made it impossible to tell at which moment they were applied (e.g. "ERPs were averaged offline over an epoch of X ms time-locked to the onset of the target, with a Y-ms prestimulus baseline. Trials contaminated by eye movements or blinks (electrooculogram > Z mV) were rejected.", or "Averages of artifact-free ERP trials were calculated for each experimental situation after subtraction of the 100 ms pre-stimulus baseline").

There were two more issues, which exceed the scope of this review due to complexity of their quantifying, but they are also worth considering. First, as it has been mentioned, in papers in which the order of operations could be assumed, there were additional cases in which it could not be confirmed with certainty. Second, not all orders of processing steps were optimal, or in some cases acceptable, as it can be seen from the earlier described example of a paper in which high-pass filtering was applied after averaging.

Together with information on insufficient reporting on individual processing steps, this result shows that improvements in the way ERP data processing is typically presented are necessary in order to allow adequate evaluation of a study, as well as replication attempts based on articles.

[16] Re-referencing was not the only processing step that was reported in a section other than the one on pre-processing – some steps were described in sections describing measurement and analysis, or even results, making reconstructing the order of operations harder.

N400 amplitude measurement and statistical analysis

ERP measurements result in an abundancy of data. Consequently, if a priori hypotheses are not specific and exploratory approach is not followed by proper correction for multiple measurements, Type I error probability easily becomes so large that it approaches certainty. There are two practices which contribute to Type I error rate increase, and which are not always obvious to researchers.

The first issue is circularity – using the same dataset for selecting a subset of data to be compared in the next step (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). The most extreme example of this issue is visual inspection (see Luck & Gaspelin, 2017), but there are other strategies susceptible to this pitfall, as well. For example, another strategy used in ERP research is to compare two waveforms by making a series of comparisons for each time point, and then to apply another test to the same waveforms, this time to compare the time windows which encompass time points shown to differ in the first step.

The second issue is not correcting for multiple comparisons of main effects and interactions (Luck & Gaspelin, 2017). These are typically treated as planned comparisons, but such an approach is not appropriate when there are many effects and interactions.

While these problems have always been inherent to ERP research, in its early period, there were not many options for data analysis or previous studies to provide grounds for specific hypotheses. The problem used to be less urgent, too, at least when it came to electrode choice, as there were typically only a few recording sites. However, the circumstances have changed and more recent guidelines (Picton et al., 2000) first stressed the importance of setting specific hypotheses in advance, and now the most up-to-date standards (Keil et al., 2014) require authors to justify their choice of measurement windows and electrode sites and to ensure that it is not biased towards finding a statistically significant effect. Likewise, the readers are reminded that appropriate adjustments for Type I error rate must be made.

Amplitude measurement.

Grounds for choosing analysis window.

To see how much this issue was given attention in the reviewed literature, we investigated several aspects of the reviewed studies. In this section, we will focus on strategies employed to define the N400 measurement window. All strategies were grouped into eight basic approaches, some with subgroups, and additional details were noted when relevant (see Table 2.2). In one study, its two experiments had measures based on different approaches, and there were 11 papers in which the researchers listed multiple arguments that they considered in order to select their analysis window. These papers are included in percentages of all approaches they included. When multiple arguments were given for the same measurement window, the researchers usually did not report how they combined them to reach the final decision when the strategies they used did not result in the same conclusion.

Table 2.2. Frequencies of strategies to select latency window for the N400 amplitude measurement

Reason for choosing measurement window	**f**	**%**
justification not given	45	34.09%
visual inspection	34	25.76%
visual inspection of raw waveforms	*32*	*24.24%*
visual inspection of GFP graph	*2*	*1.52%*
cite previous research	23	17.42%
the same window was used as in the cited paper(s)	*12*	*9.09%*
the same window not found in the cited paper(s)	*11*	*8.33%*
data analysis algorithms to locate the effect (PCA, TANOVA, cluster-based permutation test)	4	3.03%
statistical analysis on multiple consecutive or overlapping shorter windows	3	2.27%
consecutive significant comparisons – each time point or short window analyzed separately with the goal to identify the measurement window	3	2.27%
strategy to identify window based on the test results given	*2*	*1.52%*
strategy to identify window based on the test results not given	*1*	*0.76%*
quote theory or own work/experience, no reference to another paper	3	2.27%
each time point analyzed separately, but not with the goal to locate the measurement window for statistical analysis	2	1.52%
window from a pilot study (choice in the pilot study not elaborated)	1	0.76%
multiple considerations for the same analysis	11	12.64%
visual inspection, quote theory or experience (references not given)	*4*	*3.03%*
visual inspection, cite a study in which the same window was not found	*2*	*1.52%*
visual inspection, cite a study in which the same window was used	*1*	*0.76%*
visual inspection, cite a study in which the same window was not found, and statistical analyses of multiple consecutive 100 ms windows	*1*	*0.76%*
cite a study, the window verified by visual inspection – the window was not found in the cited paper	*1*	*0.76%*
cite a study in which the same window is used, a series of comparisons of consecutive segments to identify the time range of the effect	*1*	*0.76%*
cite a study in which the same window was not found, statistical analyses of multiple consecutive 50 ms windows	*1*	*0.76%*
two analyses with different strategies	3	2.27%
algorithm (PCA), no justification given	*2*	*1.52%*
visual inspection, quote theory or experience (references not given)	*1*	*0.76%*
Total	**132**	**100%**

The concern with the risk of inflating Type I error rate, expressed by the authors of more recent guidelines, was shown to be justified. Namely, approximately a third of all reports (35.61%) did not include any reasoning behind choosing one analysis window over its alternatives. When papers did include N400 window choice justification, the most common strategy was visual inspection (33.33%). The researchers reported using only visual inspection to select their analysis window in 35 papers (26.52%), and it was one of the arguments in the remaining nine papers (6.82%). In eight out of nine of these papers, the researchers cited either previous research in general, without referencing a specific paper, or they referenced papers in which a different N400 window was used. Therefore, visual inspection was the deciding factor in these cases, too. In one paper, the window was selected based on another study and the data were inspected visually, but the window was not changed after inspection. In most cases, the researcher inspected raw waveforms, but in two studies, the graph showing global field potential was used instead.

The second most frequent strategy was to use the window that had been used in an earlier study or reported in a review paper (30 papers, 22.73%). In 23 of these papers, the researchers did not report considering any other arguments. In 6 papers, additional, data-driven, window choice strategies were explored, and in one paper, the authors reported that they confirmed that the selected window was appropriate using visual inspection. However, a more careful examination revealed a concerning trend. Namely, when we examined the literature cited in these papers, in case of 11 of the 23 papers from the first group, we found that the authors had cited literature in which a different measurement window was taken. This was also true of the paper in which the authors reported that they had only confirmed the selected window using visual inspection, as well as four out of six papers in which the authors listed other arguments in addition to the cited papers. Taken all together, more than a half (16 out of 30) of papers that cited another paper as justification of their measurement window choice, referred to literature that did not support this choice. This shows that the authors did not rely on these papers to reach their final decision and that their approach was, in fact, data-dependent, and, consequently, that their report was not accurate, as it misrepresented their strategy.

Other strategies were not frequent. Authors of seven papers referred to theory or their own experience in general, without providing a specific reference. In four of these seven papers, the authors also used visual inspection, as mentioned earlier, so the previous research was a general guide to determine if an effect was indeed the N400, rather than a basis to determine precise boundaries of the measurement window.

In four studies, the dependent variable was a factor score derived from the principal component analysis (PCA), instead of an amplitude measure based on a window. In two of these studies, in addition to analyses of PCA factor scores, a traditional window-based measure was used, and it was not reported how this window was selected. In one study, a cluster-based permutation test was performed – all time points and electrode sites were analyzed separately to locate regions and time windows significant after correction for Type I error.

Topographic analysis of variance (TANOVA) was used in one study to identify a window that was analyzed using the traditional, window-based approach.

In three papers, a series of comparisons was conducted separately for all time points or very short windows (e.g. 10-20 ms). In these papers, the effects that were significant for at least a set number of consecutive windows were considered reliable, and they were subjected to the window-based statistical analyses. As Luck (Luck, 2014) explains, the issue with this approach is how to properly determine the number of consecutive significant comparisons necessary, and the papers did not elaborate how this problem was solved. In two of three papers, it was reported how many consecutive time points should be significant for an effect to be deemed reliable, and this information was omitted from one paper.

Similarly, authors of four studies performed statistical analyses on short adjacent time windows covering the entire time range and reported which time periods differed significantly. The difference from the previous approach was that the analysis windows were longer (25 ms and 100 ms), there was no condition that a certain number of consecutive windows needed to be significant in order for the effect to be considered reliable, and the analyses were conducted separately for each window – the approach was not used to define a time range that would be subjected to the main analysis as a whole. In three studies, the windows were consecutive, and in one study they overlapped (50 ms windows were analyzed, and each window started 25 ms after the previous one). Given that there was no statistical correction for multiple comparisons in these studies, this approach is prone to Type I error. On contrary, in the study with overlapping windows, the time ranges in which the p value was close to the $\alpha = 0.05$ threshold were also presented as the windows in which an effect was found, making Type I error almost certain.

In two papers, each time point was analyzed separately, without correction for multiple comparisons, and overall patterns of statistical significance were described. In one paper, the analyses were performed on raw waveforms, and in the second paper, on source waveforms derived from a localization analysis. Like the previous approach, this strategy is also biased towards finding statistically significant differences.

Finally, in one publication, a pilot study had been conducted, and it was used to decide on the measurement window for the main experiment.

In summary, despite decades of N400 research and its relatively stable latency range, most studies took exploratory approach to determining the measurement window. Two concerning trends were observed. First, in most cases, exploratory strategies were in most cases biased towards finding statistically significant effects. Second, when previous research was used to justify measurement window selection, it was frequently found that the authors' choice was not supported by the cited literature, even though this was implied in the article. Additionally, when multiple arguments were given, it was not explained how the final decision was made given that different strategies rarely produce identical results.

Latency range.

The latency range used for the N400 measurement was reported almost universally – it was omitted from only one paper.

In 123 papers, the N400 effect was either measured based on a set window, or a time range of the effect was reported as the outcome of the main statistical analysis (e.g. cluster-based permutation testing). These papers included reports on 69 different N400 measurement window choices. The most commonly reported measurement window was 300–500 ms (19 papers, 15.44% of the studies in which a latency range was reported), followed by 350–500 ms (10 papers, 8.13%). Other windows appeared in fewer than 5% of all papers, and 53 of them appeared only in one study. The broadest analyzed time range was 200–800 ms, which was divided into six 100-ms measurement windows. This range also had the earliest starting and the latest end point. The longest duration of a single measurement window was 400 ms (300–750 ms), and the shortest window lasted only 24 ms (406–430 ms).

In 20 papers (16.26%), there was more than one measurement window. In fourteen cases (11.38%), the time range was divided into two or more, up to nine, shorter measurement windows. In the remaining six papers, different windows were used for different purposes – different experiments, types of stimuli, separately analyzed anterior and posterior effects, and to test the results of two different strategies of window selection (based on visual inspection and previous literature).

Which window should a researcher choose if they wanted to base their decision on the existing literature? Given the described variability in measurement window choices, we examined the entire epoch to locate the time range in which they overlapped.

For this analysis, information on the measurement window was extracted for all separate datasets reported in the reviewed literature. If a paper included more than one experiment with the same sample, this was considered one dataset, and multiple experiments with separate samples from the same publication counted as separate datasets. If multiple measurement windows were used to analyze the data obtained from the same sample or experiment, they were represented by one window, whose bounds were the earliest and latest points in all measurements. Papers that did not include information on measurement or sample size were excluded from this analysis. As a result, latency ranges for a total of 133 experiments were extracted from 120 papers which had information on both sample size and N400 latency range.

Next, the information on whether each time point was used for the N400 measurement was extracted separately for each dataset. The 0–850 ms epoch was selected for this purpose because it encompassed latency ranges from all experiments. In addition, time ranges from each dataset were weighted by the number of participants per condition. Thus, each millisecond in the 0–850 ms post-stimulus epoch received a score based on the number of times it fell within the N400 range and the number of participants per group in the experiments in which it was found.

The results showed that there was a sudden drop in scores after 500 ms, and that there were two large increases – after 300 and 350 ms. The increase following 300-ms point was slightly larger compared to 350 ms, and 300–500 ms was also the most frequently used measurement window. Therefore, the researchers that would want to select their N400 measurement window a priori based on the existing literature, should use 300–500 ms window, at least in case of experiments with pictures as target stimuli. Graph 2.6 (p. 64) shows all latency ranges that were used for the N400 measurement and analysis in the reviewed literature, and its heat bar is a visual representation of the weighted frequencies for all time points. Appendix H contains a more detailed description of this analysis, while an Excel version of Graph 2.6 with all scores for the heat bar can be found in the Online supplement (Appendix A).

In eight studies, the researchers did not define any measurement windows (e.g. a PCA factor score was used as a dependent variable, or peaks were measured but they were not identified within a specified window).

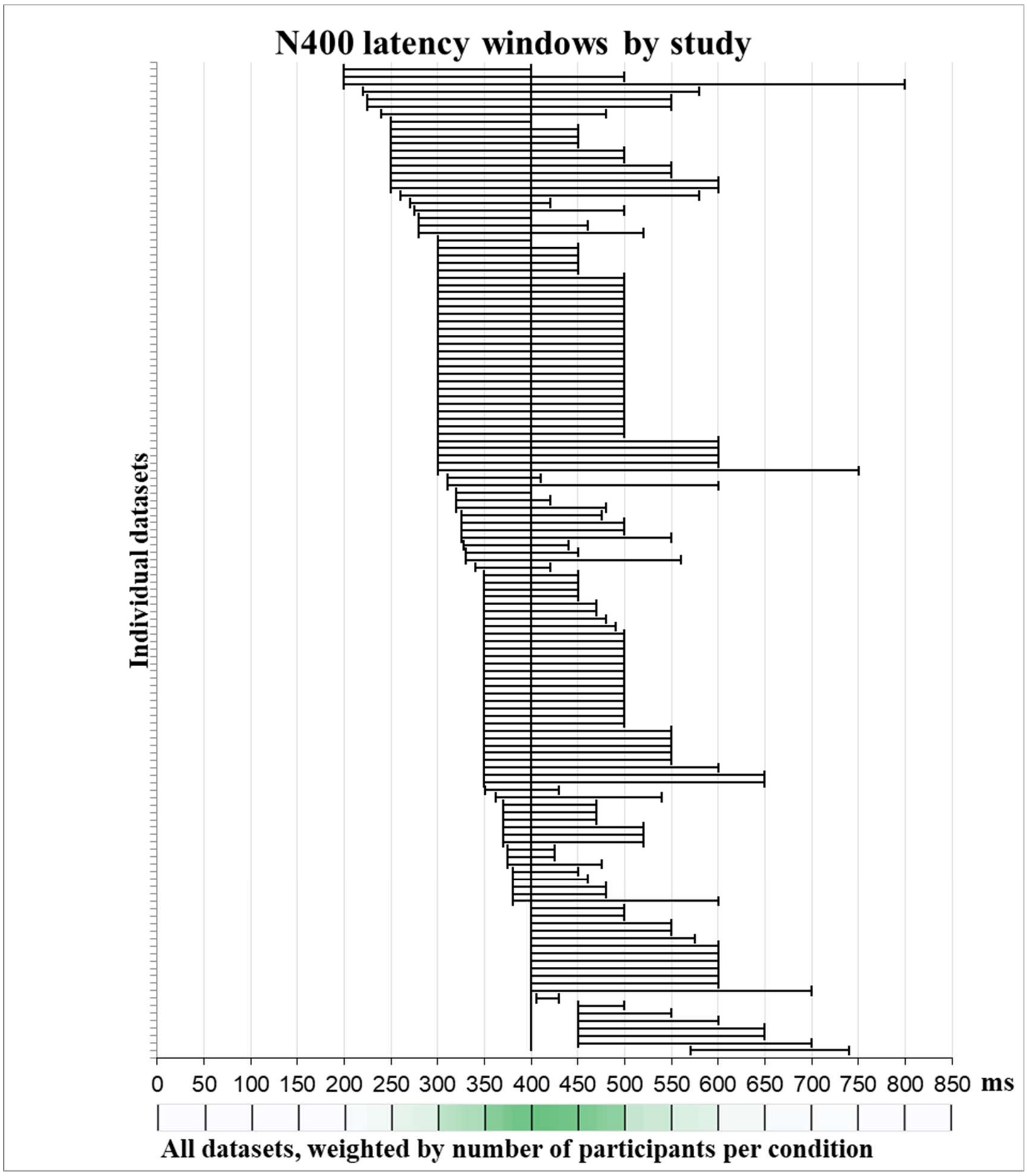
Two papers that differed from others in their approach were found in this group. In one study, the N400 was not quantified. Instead, the researchers used visual inspection to conclude that there was no N400 effect in the experiment in which stimuli were pictures. This study was included in our sample because the N400 was examined, a conclusion about presence of the effect was made and this result was interpreted together with other findings based on quantitative measurements. The study demonstrates the scope of variability in the ERP methodology that can be found in the peer-reviewed literature, and it is an example that demonstrates that visual inspection is sometimes used not only to locate an a priori chosen component, but to examine the entire waveform and determine which components will be measured. The difference between this study and other studies with visual component selection was that the researchers in other studies explicitly addressed only the waveform segments in which the differences were visually prominent and statistically significant, and it was only implied that an effect was not found in other sections.

The second paper that could be considered an outlier also speaks to the variability in how the N400 is analyzed, but also a variability in what is labeled as N400. In this study, the N400 was identified in response to the first stimulus in a task involving learning matching stimuli pairs. This component, therefore, could not be the same as the N400 found in other studies. However, given that the effect was labeled and interpreted as the N400, the study was not excluded from this review. Inclusion of this study did not affect our analyses of the N400 localization and timing, because it was not included in either of these analyses.[17] More generally, this study is one of the more extreme examples of heterogeneity of study designs that the researchers used to produce and measure effects considered to be related to the N400 in the reviewed literature.

Amplitude measure.

Information on the type of measure used to quantify N400 amplitude could be extracted from 123 papers (93.18%). Seven publications (5.30%) did not provide this information and two (1.51%) were marked as inconclusive.

[17] Amplitude in this study was measured using a peak which was not defined with respect to a specific window, and a broad set of electrodes was used, so this study did not meet the criterion for the inclusion in the subset of papers used to identify the most frequently analyzed electrodes.



Graph 2.6. N400 window choices in all datasets, i.e. experiments on separate participant groups, in papers in which an N400 analysis window was reported (N = 133 datasets from 120 papers from which both sample size and latency window could be extracted). If a paper reported multiple analysis windows or multiple experiments on the same subjects, it was represented by a single window, whose lower and upper bounds were the most extreme measures of all windows reported in this paper. Bands show N400 latency ranges for all individual datasets. The heat bar in the bottom displays frequency of including each time point (1 ms) in the N400 latency range, weighted by the number of participants per condition for each dataset. Shades of green show differences between the lowest (white) and the maximum weighted frequency (dark green). The color scale was created in Microsoft Office 365 Excel (version 1909, www.microsoft.com).

Most papers had one type of measure for the N400 amplitude. More than one amplitude measure was reported in 4 papers, 3.25% of the total number of papers from which the amplitude measure could be extracted.

The most frequently reported measure was mean of the measurement window, which was used in 74.80% papers. In addition to reporting mean amplitudes, peak amplitude was measured in one paper, and mean amplitude was the only measure in the remaining publications.

Peak amplitude was the second most frequent measure. It was recorded in 18 studies (14.63%). Despite the advantages of isolating a component using difference waves (Luck, 2014), it was most commonly measured from single waveforms (15 papers). However, three papers reported peaks measured from difference waveforms, and peaks were measured both from single and difference waves in one paper.

Other amplitude measures were not frequent. Mean area within a window was used in seven studies (5.69%). PCA factor score in four papers (3.25%). Mean amplitude of a short window centered on a peak was used in three papers, and in each it was measured in a different way. The windows included 20 or 50 ms on each side of the peak in two studies, and 20 ms on each side of the peak in global field power charts in the third study. Several approaches were unique to papers in which they appeared: analysis method developed by Hoorman, Falkenstein, Schwarzenau, & Hohnsbein (1998), analysis of each time point separately in ERP grand averages and in source waveforms obtained from dipole analysis, and the earlier described decision not to quantify the N400 based on visual inspection. Notably, in one of the two studies marked as inconclusive, cluster-based permutation testing was also used to analyze individual time points, but this approach was marked as inconclusive because the authors also reported using mean area measures of amplitude, which cannot be calculated from individual time points.

Main statistical analysis of the N400 amplitude.

Grounds for choosing electrode locations.

Choosing analysis window is not the only analysis decision with many researcher degrees of freedom – EEG recordings typically involve dozens of electrode sites, which can be grouped in hundreds of different ways, so we investigated the arguments for selecting one set of electrodes over others in the main statistical analysis of the N400 amplitude.

Out of 132 papers included in this review, it was not possible to extract information on which electrode set was used from five papers (3.79%), so the grounds for their choice were not examined. This leaves 127 papers that were summarized here.

The papers could be categorized into nine groups based on their arguments for electrodes selection and groping. Frequencies and percentages of each category are presented in Table 2.3.

Table 2.3. Frequencies of strategies to choose analysis montage within papers from which information on analysis montage could be extracted

Reason for choosing analysis montage	**f**	**%**
justification not given	48	37.80%
all electrodes used in the analysis and analyzed separately	19	14.96%
visual inspection of voltage information	17	12.88%
previous research: montage from a cited paper	12	9.45%
cited a paper, but do not use the same montage	12	9.45%
automatic procedures for clustering data (PCA, cluster-based permutation test)	4	3.15%
quote theory or own work/experience, no reference	2	1.57%
no choice – only one electrode site recorded	1	0.79%
visual inspection of current source density maps	1	0.79%
smallest t value/largest effect size	1	0.79%
multiple arguments	8	6.30%
visual inspection + cited a paper, but use a different montage	*3*	*2.36%*
visual inspection + quote theory or own work/experience, no reference	*3*	*2.36%*
visual inspection of voltage information and t statistics maps + cite a paper, but use a different montage	*1*	*0.79%*
quote theory, no reference + smallest t value/largest effect size – part of electrodes chosen by the first criterion and part by the second	*1*	*0.79%*
two different models	2	1.57%
(1) all electrodes analyzed separately and (2) smallest t value/largest effect size	*1*	*0.79%*
(1) reason not reported and (2) visual inspection	*1*	*0.79%*
Total	**127**	**100%**

Note: Five papers were excluded from this list because it was inconclusive or not reported which sites were analyzed.

Most papers included one main analysis, or a set of main analyses with different combinations of experimental factors, but the same electrode model, so they fit only one of these categories. However, there were two papers in which analyses of main experimental factors included two different electrode choices. In one of these papers, the choice of one analysis model was not justified, and the second model was based on visual inspection. In the second paper, one analysis focused on the location of the largest effect size/p value, and the second analysis included all recorded channels as separate levels of one electrode site factor. Additionally, authors of 10 papers reported considering multiple different arguments when deciding on the analysis montage. Frequencies of electrode selection strategies presented below will include these papers.

Like in case of measurement window selection, more than a third of papers (38.58%) did not report the reason for choosing analysis montage. The second most common group (15.75%) were studies in which all recorded sites were analyzed individually, without grouping them into regions or organizing them into multiple electrode site factors.

Among studies in which electrodes were either grouped or a subset of them was selected for analysis, the most frequently reported basis to achieve this was visual inspection, which was used in 19 (14.96%) papers. In addition to these 19 papers, the researchers relied on multiple arguments when deciding on the analysis montage in eight studies (6.30%). The decision was based on a combination of visual inspection and previous research in seven of these eight papers. In some cases, visual inspection was combined with a reference to theory or previous research body in general, without pointing to a specific paper, and, in others, the researchers cited earlier publications, but they did not use the same montage as the cited literature. In other words, visual inspection was the deciding factor for selecting the exact analysis montage, while cited studies were used either to inform expectations about the approximate region in which an N400 effect could be expected, or to confirm that the registered effect was indeed the N400. Taken together, this amounts to 26 (20.47%) papers in which visual inspection was the sole or the main deciding factor when choosing analysis montage. In most papers in which visual inspection was used, the researchers inspected raw ERP waveforms. CSD maps were inspected in one paper, and both raw data and t statistic maps were used to select electrodes in another paper.

In addition to visual inspection, researchers frequently cited previous research as basis on which they made selection of the analysis model. In 12 papers (9.38%), the researchers used the same analysis model as the cited study. In another 16 (12.50%) papers, changes were made to the model from the cited study. Four of these were papers in which previous research was used to constrain visual inspection, as described in the previous paragraph. The remaining 12 papers included studies in which researchers reported using a model similar to a previous paper, and papers in which the researchers cited a paper that did not support their choice of analysis model, a case comparable to the earlier described issue of mismatch of analysis window choice between the analyzed and cited papers.

There were other approaches, as well, but they were less frequent (< 5%): reference to previous research body without providing individual references, automated algorithms based on techniques such as cluster-based permutation test or principal component analysis to locate effects, selection of sites of the largest effect or the smallest p value, and recording only one electrode and consequently avoiding the question of electrode choice. Additionally, in the one remaining paper out of eight in which multiple arguments for choosing the analysis model were listed, two pairs of electrodes were analyzed, and each pair was selected using different criteria.

In summary, similarly to window selection, many papers did not include justification for selecting and organizing electrodes into factors. Among those that did, three common approaches were registered: analyzing all recorded electrodes without grouping them, relying on visual inspection or using previous research to guide decision process. When the previous research was used to inform decisions on analysis montage, in most cases the researchers did not use the same model as the studies they cited, leaving many researcher degrees of freedom in choosing their own model. Exploratory approaches were often biased towards finding statistically significant differences – visual inspection, selecting location with the largest effect, but it was not always the case – in some studies, all recorded sites were organized as separate levels of one factor, or used as basis for a PCA or cluster-based permutation testing.

Other approaches recommended in literature (Keil et al., 2014; Luck, 2005, 2014; Luck & Gaspelin, 2017) were not found in the reviewed papers.

Which sites were chosen for the main analysis?

To investigate analysis montage, we asked three questions: (1) how many electrode sites were analyzed in the main analysis in each paper, (2) whether locations were the same for papers with the same analysis montage size, and (3) which electrodes sites appeared most frequently in analysis montages.

Information on analysis montage size could be extracted from 122 papers (92.42%). In 121 of these papers, the N400 effect was quantified, and in one paper, the conclusion about its absence was made based on visual inspection instead. Ten papers (7.58%) were labeled as inconclusive due to missing or conflicting information. The results demonstrated that the analysis montage size varied a lot in the 121 papers from which this information could be extracted. There were 41 different electrode counts, 18 of which appeared in only one paper. They ranged between 1-144 sites (M = 22.08, SD = 25.43). Five most frequent electrode counts were 6 electrodes (9.02% conclusive papers), 12 electrodes (8.20%), 29 electrodes (6.56%) and 9 and 26 electrodes (5.74% each). All other analysis montage sizes appeared in fewer than 5% of all papers. In some of these studies, the researchers included all sites that they had recorded, while in others, a selection of recorded locations was made.

Based on the analysis montage size, it can already be concluded that there was not much consistency in which electrode site combinations were chosen for the analysis. In addition to the 10 papers from which information on the analysis montage could not be extracted, there were 5 studies (3.79% of all publications) in which all electrodes were included in the main statistical analysis, but it was not reported which locations were recorded. As a result, analysis montages could be extracted from 117 papers (88.64% of the total sample). These 117 papers could be divided into 93 separate groups with distinct electrode choices, as there were up to 10 unique combinations of electrodes for a given analysis montage size. Most of these electrode choices, 77 of them, appeared in only one publication, and the remaining 15 montages appeared in 2-4 different papers. Three analysis montages appeared in four publications: (1) only Cz, (2) F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4, and (3) a montage of 29 electrodes that has been described in the section on recording montage (e.g. Eddy et al., 2006). The first two montages were used by different author groups, but the third was used in four related studies which were all conducted by the same group of authors.

While there was a lot of variability in which electrode combinations were analyzed, there was some overlap in which electrodes were part of these combinations. In order to provide guidance for deciding on the analysis montage based on previous literature, we examined which electrodes were reported in studies in which up to 12 electrode sites were analyzed. As explained in the Codebook (Appendix D), this cut-off point was chosen because montages with more than 12 electrode sites typically involved analyzing all or most of the recorded sites, which distributed over the entire scalp, while the smaller recording and analysis montages were more frequently restrictive.

For this purpose, data on 65 experiments conducted on different samples was extracted from 58 publications. Within analysis montages used in these experiments, 66 different channels were found. Frequency of using each channel for analyzing data from the selected 65 experiments was registered, and, additionally, this information was weighted by the number of participants per group in each of the 65 experiments. All electrodes used in the analyses are shown in Figure 2.2, in which weighted frequency of each site is presented using color scale. More information on this analysis can be found in Appendix H.

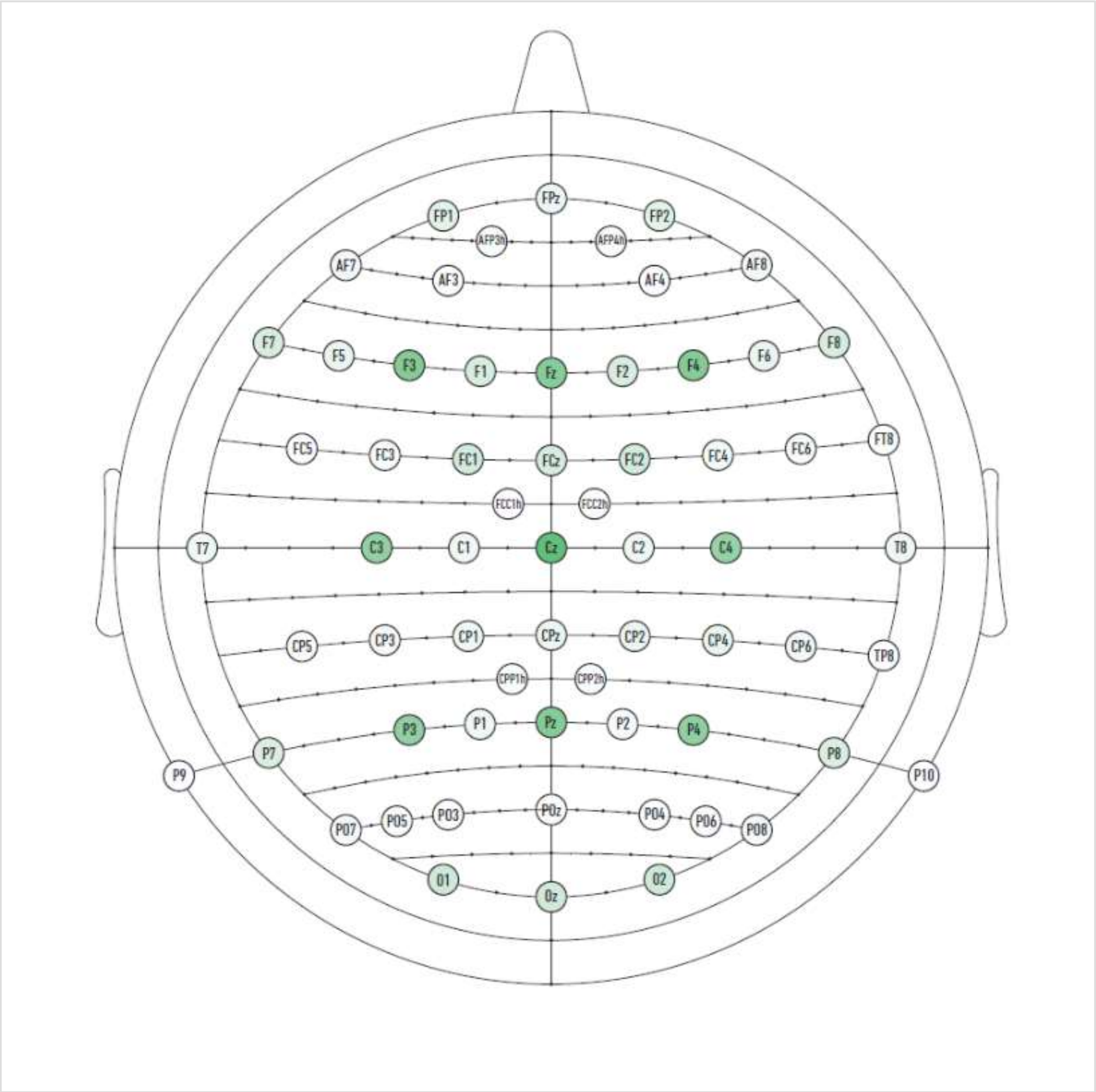


Figure 2.2. The montage shows all electrodes that were used for measurement of the N400 in the main statistical analysis. Only studies with 12 or fewer electrodes were used to generate this montage, because larger montages more frequently included analyses of the entire scalp with broadly distributed electrodes. If a paper included more than one experiment with different subjects, both experiments were included in the analysis separately. Shades of green show differences between the lowest (white) and the maximum frequency (dark green) of using an electrode, weighted by the number of participants per condition for each experiment.

Nine electrodes stood out compared to others: F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4. Each of these electrodes were used in 23 or more experiments, compared to all other sites, which were included in analyses of 10 or fewer experiments. The results were the same when data was weighted by the number of participants per group, as it can be seen in Figure 2.2. Cz was the electrode most commonly used for the N400 measurement, compared to the other eight sites. It was included in analyses of data from 36 selected experiments (55.38%). In other words, even the most commonly analyzed electrode appeared in a little more than a half of the selected analyses.

The described variability can be partly attributed to differences in the recording montage, but not entirely, given that even different montages frequently overlapped on many electrode sites, and that part of the variability in recording site montages could be reduced if there was more consistency in which locations from the same caps were chosen for recording. The variability was more likely the consequence of the method of electrode location selection, described in the previous section. The large number of researcher degrees of freedom involved in many decisions on the electrode location and the frequency of analyzing all electrode sites are contrasted by a much smaller proportion of studies in which the decision to select only some locations was described as a priori.

Analysis.

While the presented variability in selection of electrodes for the N400 measurement was large, these choices can be divided even further, since the same set of electrodes can be analyzed using different statistical approaches and divided into spatial factors in different ways. As a result, in the 117 papers in which the analysis montage was known, we registered 99 different ways to analyze data with respect to electrode factors. Two of these analysis strategies were found in 4 papers – one strategy was to analyze only Cz, and the other was used in the four related studies by the same author team, described in the previous section. Two more analysis strategies were implemented in three papers each, eight strategies in two papers each, and there were 88 analysis strategies that were used only once in our sample.

These diverse analysis strategies could be grouped into categories based on which statistical tests were used and how the electrodes were organized into factors. Including papers in which the analysis montage was not given, 92.42% of all papers (f = 122) could be categorized into broader groups according to their analysis strategy. On top of the 10 papers that could not be categorized, there were three which were included in the following analysis, but some details were marked as inconclusive: PCA on a montage that was not provided, ANOVA on analysis windows that were not reported, and inconclusive experimental factors.

Analysis of variance (ANOVA) was used by most studies. It was used in 112 publications (91.80% of papers from which this information could be extracted). It was the only main analysis strategy in 110 of these papers, and in two additional papers, pairwise comparisons using Wilcoxon or t test were also used. In most papers, it was applied to the N400 amplitude measures derived from original waveforms, but in four studies, PCA factor scores were dependent variables. In all cases but one, general linear ANOVA was used, and in one study, it was linear mixed ANOVA. In the majority of studies that used ANOVA (69 papers, 61.61%), electrode factors were arranged so that individual electrode sites served as factor levels, but it was also common to group electrodes into regions of interest (ROIs, 43 papers, 38.39%). ANOVA models on single electrodes were almost as likely to include all locations that were recorded (35 papers) as to be restricted only to a selection of sites (36 papers). On the other hand, regions of interest were much more often limited to a subset of electrodes (37 out of the 43 papers in which ANOVA models were based on ROIs).

Regarding the way electrodes were grouped, the most frequent ANOVA model had one electrode site factor in which individual electrodes or ROIs were factor levels (36 papers, 32.14% papers in which the main analysis was based on ANOVA). This was the most common analysis model both for analyses of single electrodes (25 papers out of 69) and ROIs (11 papers out of 43). Analyses based on ANOVA on single electrodes also often included multiple ANOVAs, separately for the midline and lateral locations (16 papers), as well as one ANOVA with a 2x3 electrode arrangement (6 papers). On the other hand, analyses of regions usually were not separated into multiple ANOVAs, and the electrode arrangements were more frequently 3x3 (6 papers) or 2x2 (5 papers). Other ANOVA models were less frequent and appeared only in a few studies each.

Regarding analysis window, one measurement window was used in most studies (see Amplitude measurement). However, in fourteen cases, the latency range was divided into shorter measurement windows, and the number of windows varied between two and nine. This approach would be appropriate if an analysis in which all windows were levels of one factor was conducted to establish that there was an effect before analyzing each window separately, and if the appropriate corrections for multiple comparisons were used (Keil et al., 2014). However, this was not the case in the reviewed papers, resulting in Type I error rate inflation.

Other approaches were less frequent. In nine papers, paired tests were used, although, as mentioned above, in two of these studies, the main analysis strategy also included ANOVA models. In these papers, a variety of strategies was registered. They differed in the test that was applied – Wilcoxon or t test. Furthermore, in some papers, a series of paired comparisons on each time point was made, and in others the tests were applied to larger time windows. Likewise, the test could be applied to regions of interest, all individual electrodes or a subset of them, as well as to original or source waveforms, with or without appropriate corrections for multiple comparisons. In two papers, the researchers used multivariate analyses of variance (MANOVAs), and in one study, the N400 absence was determined based on visual inspection.

The frequencies of all analysis strategies can be found in Table 2.4. Due to high variability, some of the frequencies may be affected by preferences of a laboratory or team which is represented in the sample by multiple publications.

Table 2.4. Frequency of strategies for the main statistical analysis of the N400 effect.

General strategy	**Regions of interest**			**Individual sites**			**Total**
	Selected	All	**Total**	Selected	All	Total	
ANOVA							

General strategy	Regions of interest			Individual sites			Total
(1) two electrode factors: 2x2	5		**5**	2		**2**	**7**
(2) two electrode factors: 2x3	1	1	**2**	6		**6**	**8**
(3) two electrode factors: 3x3	4	2	**6**	3		**3**	**9**
(4) ANOVA on PCA scores					3	**3**	**3**
(5) one electrode factor: individual ROIs/sites	10	1	**11**	10	15	**25**	**36**
(6) separate ANOVAs for midline and lateral sites	3	1	**4**	3	13*	**16**	**20**
(7) one ROI/site – no electrode factors	4		**4**	3	1	**4**	**8**
(8) all ROIs/sites in the same ANOVA – other models	4		**4**	5		**5**	**9**
(9) multiple ANOVAs – other	4	1	**5**	2		**2**	**7**
other analyses							
(10) MANOVA				1	1	**2**	**2**
(11) approaches based on t test or Wilcoxon	2		**2**		5	**5**	**7**
visual inspection of the effect			/			/	**1**
multiple different analyses							
11 and 9	1		**1**				**1**
11 and 5				1		**1**	**1**
4 and 5					1	**1**	**1**
6 and 7	1		**1**				**1**
8 and 4					1	**1**	**1**
Total	**39**	**6**	**45**	**36**	**40**	**76**	**122**

Note: This analysis included 122 papers from which the necessary information could be extracted. * 7/13 analyses are from papers which share one or more authors.

Additional analyses of the N400 component.

On top of the main statistical analysis of the N400 amplitude, many studies had additional analyses of this component. A subset of 127 papers will be presented here. The remaining five publications were marked as inconclusive because at least some of the analyses could not be categorized due to unclear descriptions or conflicting information in methods and results sections. Overall, full descriptions of the additional analyses had some inconclusive details in 23 papers (17.42% of the total sample, 19.01% of studies with additional analyses). However, all but the five above-mentioned publications had analysis descriptions that allowed their categorization into broader groups that will be presented here.

As it can be seen in the Codebook (Appendix D), additional analyses were grouped into 12 categories: (1) post hoc and planned pairwise comparisons (with subcategories for different corrections for multiple comparisons), (2) uncorrected ANOVA post hoc comparisons of simple effects to explore interactions found in the main factorial ANOVA, (3) corrected ANOVA post hoc comparisons, (4) ANOVAs on normalized data (McCarthy & Wood, 1985), (5) ANOVAs in which the main window was divided into shorter sections to explore the time course of the event, (6) other additional ANOVAs, (7) correlations with behavioral and other non-ERP variables, (8) Shapiro-Wilk test of distribution normality, (9) analyses of effects on peak or onset latency, (10) comparisons of the N400 with other ERP components, (11) ANCOVA, (12) other analyses. The results are shown in Table 2.5.

Table 2.5. Frequencies of additional statistical analyses

Additional analyses	**f**	**% of parent row**
post hoc and planned pairwise comparisons	77	60.63 %
post hoc and planned - no correction reported	*33*	*42.86%*
Dunnet	*1*	*1.30%*
Tukey HSD	*14**	*18.18%*
Bonferroni	*20*	*25.97%*
polynomial contrasts	*1*	*1.30%*
Newman-Keuls	*4*	*5.19%*
Fisher's LSD	*4*	*5.19%*
FDR	*1*	*1.30%*
Duncan multiple range test	*1*	*1.30%*
post hoc ANOVAs – simple effects from the main ANOVA	27	21.26%
not corrected for multiple comparisons	*26*	*96.30%*
corrected for multiple comparisons	*1*	*3.70%*
other additional ANOVAs on amplitude	31	24.41%
ANOVA(s) on normalized data	21	16.54%
ANOVAs on a series of shorter windows made by dividing the main measurement window	8	6.30%
correlation with behavioral and other non-ERP variables	7	5.51%
other: ANCOVA, MANOVA, GLM-t test combination (LIMO), ERSP, cluster-based permutation testing, split-half reliability, linear mixed effect modelling, single-trial analysis	12	9.45%

Additional analyses	f	% of parent row
Shapiro-Wilk test of distribution normality	1	0.79%
peak or onset latency analyses	17	13.39%
comparison with other components - ANOVA, correlations to determine scalp similarity, etc.	8	6.30%

Note: Percentages are relative to 127 papers in which all analyses could be categorized, and relative to the parent row for subcategories. Many papers included multiple analyses, which were included in the percentages for each category, so the sum of all frequencies is not 127.
* Half of these 14 analyses are from papers which have the same first/last author.

Out of the 127 papers that were categorized, 116 (91.34%) had at least some additional analyses. About a half of them (64 studies) reported more than two, and up to six, different categories of additional analyses. Moreover, most additional analyses required conducting multiple comparisons.

Post hoc and planned comparisons were the most frequent (60.63%). Most of these comparisons were post hoc, and there were rarely limited to only a few selected pairs of conditions. Nevertheless, these comparisons did not include corrections for multiple comparisons in two out of five cases (42.86% of post hoc comparisons). Bonferroni (26.32%) and Tukey HSD (18.42% of post hoc comparisons, half of which were from the same laboratory) adjustments were most frequently used to correct for Type I error, but six other types of correction were registered in the literature (see Table 2.5).

Many studies had additional analyses of variance to further explore the effects of interest (63 papers, 49.61 % of studies that were categorized). The most common additional ANOVAs were post hoc comparisons used to analyze interactions. They were reported in 27 papers, or 21.26% of studies that were categorized. In only one of these studies, Bonferroni correction was used to control for Type I error, and in all other cases, there was no correction, even though studies typically involved multiple factors, often with more than two levels. Other frequently used approaches included ANOVA(s) on normalized data (21 papers, 16.54%) and dividing the main measurement window into shorter sections that were separately analyzed (8 papers, 6.30%).

In some papers, variables other than the N400 amplitude were included: analyses of the N400 latency (13.39%), comparisons with other ERP components (6.30%), and correlation between the N400 amplitude and behavioral and other non-ERP measures (5.51%), and comparisons with other ERP components. Other analyses were used only in a few studies.

Correction for Type I error rate.

As shown in the preceding sections, the authors of the reviewed papers used a variety of different analyses and relied often on decisions made a posteriori. Even if data-dependent selection of statistical tests is left aside, in many papers, a considerable number of comparisons was explicitly made. In some studies, the main analysis comprised multiple separate ANOVAs for different experimental conditions, electrode sets or measurement windows. In others, there were additional analyses on top of the main comparisons.

With this many effects to test, it is not appropriate to assume that all tests for main effects are a priori and do not require correction for Type I error (Luck & Gaspelin, 2017). However, this was the case in most of the reviewed papers: 96.18% of the 131 papers in which the N400 was quantified did not take any measures to control for Type I error rate of main effects and interactions. Moreover, the threshold of statistical significance was set even higher than the conventional $\alpha = 0.05$ level in one study, to $\alpha = 0.06$, and this does not include papers in which marginally significant results were interpreted in the Discussion section as if they were statistically significant.

In total, there were five studies in which there was an attempt to lower Type I error rate. In only one of these studies, the correction applied to all comparisons. Authors of this study used cluster-based permutation testing. It is an approach in which comparisons are made on each time point and electrode in order to locate data clusters which remain significant after a permutation test-based correction for multiple comparisons (Maris & Oostenveld, 2007). This strategy is one of mass univariate analysis approaches, designed for testing many individual data points while preserving more statistical power than traditional corrections for multiple comparisons (Groppe, Urbach, & Kutas, 2011).

In similar fashion, in two papers, the authors made separate comparisons for each electrode, but they selected time windows using a different approach. In one paper, authors use corrections for multiple comparisons developed by Hopf & Mangun (2000). In the other paper, two related procedures developed by the same group of authors are cited (Benjamini, Krieger, & Yekutieli, 2006; Yekutieli & Benjamini, 2001), and it is not clarified which one was used (presumably the more recent one). However, both papers included reports on multiple components, and the authors did not report using any means to correct for this fact.

Finally, authors of two less recent papers partly lowered the Type I error rate by adopting a stricter threshold for statistical significance, $\alpha = 0.01$, for some of their analyses.

To demonstrate that this issue was not negligible, we registered the number of ANOVAs, MANOVAs and ANCOVAs applied to the N400 amplitude per experiment in each study, which were not corrected for Type I error. Other analyses, such as correlations or pairwise comparisons, were not included because it was often not possible to extract information on how many comparisons were made. It is also noteworthy that most of the included analyses had several factors. In other words, the number of analyses presented here is an underestimate of the number of comparisons that were made. Additionally, we registered the number of components other than the N400 that were analyzed statistically.

There were 115 papers in which the authors both used ANOVA, MANOVA or ANCOVA and it was possible to extract the total number of these analysis. Nine studies (6.82% of total sample) did not have any of these analyses, and descriptions of the results of another eight studies (6.06%) did not allow calculating how many analyses of this kind were conducted. The remaining 115 papers included reports on between 1 and 59 analyses (M = 7.12, SD = 10.35). The largest number of analyses was registered in a paper that was not included here because the total number of analyses was inconclusive, but the authors conducted at least 576 separate, uncorrected one-way ANOVAs – one for each experimental factor, electrode site and short window.

The most frequent categories were papers with one, two and three analyses. Only one analysis was reported in almost a third of included papers (29.57%), 10.43% papers contained reports on two analyses and 9.57% papers had reports on three ANOVAs. Taken together, about a half of analyzed papers had up to three ANOVAs. The other half had four or more analyses, each with multiple factors, excluding statistical comparisons other than ANOVA, ANCOVA or MANOVA. More precisely, 33.91% papers had between 4–10 ANOVAs, and the remaining 16.52% papers had more than 10 ANOVAs. The analyses with very large numbers were usually conducted separately either on short sub-windows or individual electrodes. Granted, some of these analyses are analyses on normalized data to see if the effects remain, but the number of comparisons is still considerable.

Regarding the number of components (time windows and/or regions which were analyzed separately), we excluded three studies from the analysis because their authors analyzed the entire epoch statistically. There were between 1–14 components in addition to the N400 in the remaining papers (M = 2.63, SD = 2.58). Approximately one in nine papers did not include analyses of any other components (11.63%). Most papers included one (24.81%), two (26.36%) or three (16.28%) components in addition to the N400. About half of the 114 papers (45.61%) which involved analyses of additional components, had components both earlier and later than the N400. It was more common to test for early components (39.47% just early components, 85.08% total), than components later than the N400 (14.91% papers in which early components were not analyzed). Additionally, LRP and CNV components were measured in one paper each.

When the number of components is multiplied by the number of analyses employed to investigate them , in many of the studies included in the review it is more likely than not that some of the reported effects are a product of noise, even if the analysis window and electrodes had been chosen a priori.

Other corrections.

In the section about additional analyses, we mentioned that data normality assumption was investigated in one study. Out of other assumptions for statistical tests that were used, authors of the reviewed papers tested for sphericity, the assumption which causes the most concern in ERP analysis. It has long been known that psychophysiological data often violate this assumption because of high correlations between adjacent data points (Jennings, 1987; Jennings & Wood, 1976), and ERP guidelines (Keil et al., 2014; Picton et al., 2000) demand that researchers make appropriate corrections. Sphericity violation effects are usually mitigated by adjusting degrees of freedom by a factor of ε, using either a more conservative method, developed by Greenhouse & Geisser (1959), or a more liberal method, proposed by Huynh-Feldt (1976). Because the first method tends to be overly conservative, while the second is too liberal, some statisticians have suggested alternative strategies that combine these two corrections, such as using the average of the two correction factors (Dien & Santuzzi, 2005).

To investigate how frequently this recommendation was followed, we examined reports of 110 studies which included an ANOVA with least one factor with more than two levels. Authors of seven out of ten papers (70.00%) reported testing for sphericity and using appropriate corrections. In most cases (80.52% corrections), Greenhouse-Geisser (1959) correction was used. Huynh-Feldt method (1976) was less prevalent (18.18%). In one paper, both methods were used (details were not provided).

Topographic analyses.

On top of traditional statistical analyses, authors of 25 papers (18.94%) used spatial distribution analyses to model source waveforms or to compare component distributions. Procedures based on LORETA (Pascual-Marqui, Michel, & Lehmann, 1994) – LORETA, sLORETA, and swLORETA – were the most frequent. They were used in 16 papers (64% of all spatial analyses). Strategies derived from principal component analysis (PCA) was used in four studies, while other approaches appeared in three or fewer papers: clustering of ICA components (Onton & Makeig, 2006), spatial correlation analysis (Murray, Brunet, & Michel, 2008), TANOVA (Murray et al., 2008), LAURA (Grave de Peralta Menendez, Gonzalez Andino, Lantz, Michel, & Landis, 2001), distributed source modelling (Hauk, 2004) and Bayesian Model Averaging (Trujillo-Barreto, Aubert-Vázquez, & Valdés-Sosa, 2004). Frequencies of all analyses can be seen in Table 2.6.

Table 2.6. Frequencies of topographic analyses

Topographic analyses	**f**	**% of parent row**
all analyses	25	18.94%
LORETA	*5*	*20.00%*
sLORETA	*7*	*28.00%*
swLORETA	*4*	*16.00%*
clustering ICA components	*1*	*4.00%*
PCA	*4*	*16.00%*
spatial correlation analysis	*1*	*4.00%*
TANOVA	*2*	*8.00%*
LAURA	*1*	*4.00%*
distributed source modelling	*3*	*12.00%*
Bayesian Model Averaging	*1*	*4.00%*
partial-directed coherence analysis	*1*	*4.00%*
no topographic analyses	107	81.06%
Total	**132**	**100%**

Note: Sum of percentages of individual spatial analyses is not equal to 100%, because some publications had more than one type of analysis.

Plotting results.

The most frequent way to visualize ERPs data is, of course, in the time domain. ERP waveform graphs are so important for understanding ERP results that the demand they must be reported was made part of the first guidelines for ERP analysis (Donchin et al., 1977), and we have not registered any papers that did not show them.

One thing that is quickly observed by the newcomers in the field is that some researchers plot their ERP data with negative voltages up, and others with negative voltages down. Even Donchin et al. (1977) comment on this issue in the first guidelines, and conclude that, although it is “an eminently reasonable position” that the ERP field should standardize its polarity convention, they have not been able to reach a conclusion, because each laboratory held to their position strongly. Donchin et al. also remark that about a third of the participants at the symposium which resulted in the guidelines used the ‘positive-up’ convention.

Has either convention prevailed over time? To answer this question, we registered which one was used in each paper. Overall, the results were tied even more than at the 1974 Symposium on Cerebral Evoked Potentials in Man: 58.33% papers plotted data ‘negative-up’, 40.15% plotted results ‘positive-up’, while different conventions were used for different graphs in 1.52% cases. This review covers papers spanning a few decades, so we separately examined graphs for papers published since 2015 and until 2000 (see Trends over time) to see whether the field has been converging towards an agreement in the recent years. However, the results were essentially the same: 56% older papers had ‘negative-up’ graphs, and so did 52% of the most recent publications (excluding one paper in which both conventions were used).

In addition to the time-domain representation, many papers included some form of heat maps (48.48% publications). Voltage maps were the main solution for presenting spatial distribution (78.13% of all papers with maps), and they typically – but not always – showed a difference wave. Current source density maps were also frequently used (10.94% publications), while other types of heat maps were used only in a handful of studies (for more details, see Table 2.7).

Table 2.7. Frequencies of presenting spatial information using different types of maps

Maps	**f**	**%**
all maps	64	48.48%
voltage maps	*50*	*78.13%*
current source density maps	*7*	*10.94%*
voltage maps - normalized data	*4*	*6.25%*
statistical maps - based on t-score, p-value	*4*	*6.25%*
topographic map of PCA factors	*1*	*1.56%*
voltage maps on percent scale	*1*	*1.56%*
3D voltage maps	*1*	*1.56%*
no maps	68	51.51%

Note: Percentages are relative to the total of the parent row. Total number of publications is N = 132. The percentages do not add up to 100% because some publications contained more than one type of map.

In addition to maps and time-domain graphs, several other types of data visualizations were found, albeit each of them only in one paper: a time-frequency spectrogram, single-subject and confidence-interval ERPs (time-domain), dynamic maps *time* × *electrode* × *t* statistic and time × *electrode* × *voltage*, and a scatter plot of individual variability of components.

General considerations.

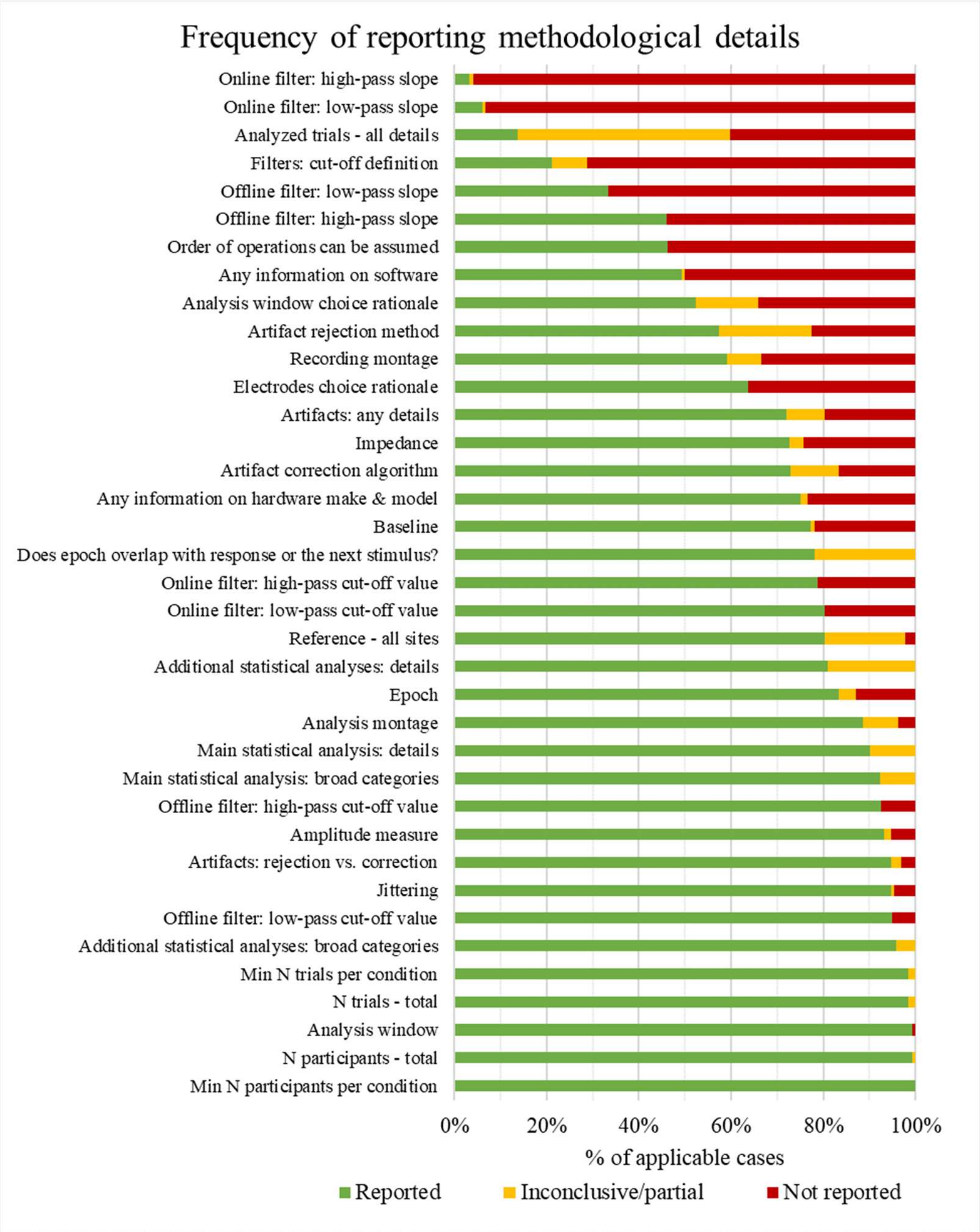
Finally, there are two general comments, concerning all statistical analyses.

First, analyses marked as inconclusive were not only sporadic. Descriptions of the main and additional statistical analyses had between 3.79 – 17.42% inconclusive details, depending on the level of detail that was extracted. There were two main reasons for labeling a study as inconclusive. In some publications, we found discrepancies between the Methods and Results sections. For example, a measure or an analysis could be described in the Methods section, but not mentioned in the Results section, or the Results section contained different analyses than the Methods section. The second reason for labeling a study as inconclusive was if some information was missing (e.g. about analysis window(s) or electrode factor(s).

The second, related issue was already mentioned in the section about Type I error rate: it was not sometimes difficult, or even impossible to determine the exact number of comparisons that were made. This problem usually occured when there were a lot of comparisons, and the authors only described statistically significant effects in the Results section, with a general remark that the other comparisons were not. This is a common practice and it is understandable because it makes the report easier to read, but it can lead to two issues. First, it doesn't put the statistically significant results within the context of how many uncorrected comparisons were made in total. Second, not knowing which exact comparisons were made is an obstacle for replication attempts. Two studies found in our sample (Butler et al., 2013; Demiral et al., 2012) show how this problem can be mitigated – in both studies, the authors provided supplementary documents in which all comparisons, both significant and not, were presented.

Reproducibility of studies based on reports

To summarize all data provided in the preceding sections, 61 papers (46.21%) were categorized as inconclusive or contained details labeled as inconclusive on variables containing verbal descriptions. When we examined different reporting categories, the ones which had the most inconclusive cells were: details of additional analyses (23 papers), main analysis montage (13 papers, on different variables), the indicator of range of trials averaged per condition (minimum of trials averaged or threshold for rejecting participants, 11 papers), rejection procedure (f = 11), and recording montage (f = 9).



Graph 2.7. Frequencies of omitting methodological details from reports. The *y* axis shows methodological information that was examined, while the *x* axis shows the percentage of papers in which this information was provided, partly provided or not provided. All percentages are relative to the number of cases relevant for the variable in question (e.g. studies in which a procedure was used). Green bars show percentage of papers in which the methodological information in question was provided. Yellow bars show percentages of papers in which some information was given, but it was either partial or inconclusive. Red bars show percentages of papers from which the detail in question was omitted. Table of frequencies and more details on them can be found in Appendix I.

In addition, at least some details were missing from all papers. The most critical variables, which were not reported by most papers were filtering properties other than cut-off (not reported in 95 or more papers, depending on the variable), indicator of range of trials averaged per condition (f = 103 papers), and equipment and software (88 or more papers; note that guidelines do not demand that this information must be provided, except for amplifier make and model). However, even when these variables were not taken into account, there were only two studies in which all other information was provided (conducted by Cansino, Hernández-Ramos, & Trejo-Morales (2012) and by Federmeier & Kutas (2002)). Other frequently omitted information included: order of operations (77 papers), mean number of averaged trials (f = 76), any information on the number of averaged trials (f = 56), rationale for electrode (f = 49) and analysis window choice (f = 47), recording montage (f = 44), artifact rejection method (f = 34).

This information is graphically summarized in Graph 2.7, where it was organized in a slightly different way. The graph shows percentages of papers in which (1) the methodological information in question was provided, (2) some information was given, but it was either partial or inconclusive, or (3) the detail in question was omitted. For more details about Graph 2.7, see Appendix I.

Aside from the report itself, very little supplementary material was identifiable through analyzed papers, even for more recent studies. Most papers (86.36%) did not refer to accessible supplementary materials other than reports on additional analyses. Admittedly, in one of them, readers were informed that data was stored on a departmental server and could be accessed by contacting authors or the department, while another paper provided a link to a Harvard Dataverse page, albeit locked to website visitors even after registration. Additionally, 9.85% papers provided only lists of stimuli descriptions, and another 2.27% provided actual stimuli or information needed to identify them in published databases of images. There were, in fact, only two papers in which access to ERP data had been provided – a link to behavioral and raw ERP data in one paper, and to component mean amplitudes in the other. There were no studies with published codes for stimuli presenting, ERP data processing or analyses.

Trends over time

As shown in Table 2.1, the oldest paper included in this review was published in 1988. Reflecting growth in ERP use, the papers are not distributed evenly. Instead, their number grew over time. Approximately a half of all papers (50.75%) were published in the last ten years, since 2010.

Over the past three decades since this publication, many things have changed in the way ERP data is collected, processed and analyzed – new technologies and analyses have become available and we have learned new things about ERP methodology. This is reflected in changes between different versions of guidelines. Conveniently, 25 papers included in this review (18.94%) were published between 1988–2000, when the first detailed guidelines for ERP research were published (Picton et al., 2000), and the same number of publications came out since 2015, a year after presenting the latest version of the guidelines (Keil et al., 2014). We present a brief comparison of these two groups, to show how improvements in ERP methodology and recommendations were reflected in practice.

Study design and sampling.

Several aspects of study design have changed over time. First, more recent studies had more participants per condition (M_{old} = 15.36, n_{new}= 18.52), even though between-group designs, which are less powerful, were more frequent in the older literature (f_{old} = 24%, f_{new}= 8%). Contemporary studies also had more trials per condition, even after excluding two studies, one in each group, which had unusually large numbers of trials per condition (M_{old} = 39.38, M_{new}= 50.74, excluding outliers). The two groups of studies did not differ a lot, however, when it comes to reporting on how many trials were averaged together – about half of papers in both groups did not report outcomes of artifact rejection, although the number was slightly higher in the sample of older papers (f_{old} = 56%, f_{new}= 44% for not reporting). Jittering interstimulus or intertrial interval became more widespread over time (f_{old} = 44%, f_{new}= 12%), while self-paced timing was more frequent in the older literature (f_{old} = 16%, f_{new}= 0%). Authors of earlier studies used tasks with delayed response and no response to the N400-eliciting stimulus equally ($f_{no\ response}$ = 20%, $f_{delayed\ response}$= 20%, $f_{neither}$= 60%), while delayed response was a preferred solution for eliminating brain activity related to motor response in the more recent studies ($f_{no\ response}$ = 8%, $f_{delayed\ response}$= 32%, $f_{neither}$= 60%).

Apparatus and software.

Equipment and software were more frequently described in more recent publications (cap reports: f_{old} = 28%, f_{new}= 76%; amplifiers reports: f_{old} = 44%, f_{new}= 76%; software reports: f_{old} = 0–20%, f_{new} = 36–68%, depending on the category). This was hardly surprising, especially for software, due to more recent development of widely available commercial and open-access software packages, as well as more complex procedures for data processing and analysis, offered by these packages.

Recording and pre-processing.

Older publications reported impedances more frequently than more recent ones (f_{old} = 80%, f_{new} = 64%). This is related to the fact that high-impedance amplifiers were often used in contemporary studies (f_{new} = 40%), but none of the authors of more dated papers reported using such equipment. As explained in the section on impedances, researchers who used high-impedance amplifiers did not provide alternative data quality indicators when impedance information was not available.

Recording montages have become bigger since the early studies. The average number of electrodes in the montage increased form M_{old} = 13.38 to M_{new} = 55.04. Montage sizes in older papers were also more diverse, while 4 out of 10 more recent studies were recorded with 62-64 active channels.

Voltage reference of choice has also changed over time. Linked mastoid or earlobe references were often used by early researchers (f_{old} = 56%), while other solutions were diverse and infrequent. In the latest studies, linked references have been abandoned for superior offline references, mean mastoids (f_{new} = 40%) and average reference (f_{new} = 28%). In case of the latter, the authors described the recording montage in only one paper.

Expansion of digital filtering tools allowed filtering data with a narrower bandpass offline. Among the older publications, five had reports on low-pass digital filters and one mentioned high-pass filtering. In contrast, data was filtered digitally in more than half of more recent studies ($f_{high-pass}$ = 56%, $f_{low-pass}$ = 64%). Online filters were described in all older publications. More recent papers, however, usually only had descriptions of analog filters when digital filters were not used. Only 3 out of 16 contemporary papers which mention digital filters also included information on analog filters. Roll-off was described by 8% older and 24% more recent papers, and it was provided for offline filters in all cases but one. Cut-off type was specified for all filters in 60% older publications, and in 12% more recent publications. Even though almost all sources (Cook & Miller, 1992; Keil et al., 2014; Luck, 2005, 2014; Picton et al., 2000) advise against notch filters, they have not been abandoned yet (f_{old} = 12%, f_{new} = 16%).

Similarly, development of better artifact correction algorithms and increased availability of programs which implement them resulted in a shift from primarily rejection (f_{old} = 88%) to combining rejection with correction (f_{old} = 32% for rejection, f_{new} = 48% for combined methods).

Baseline duration differed between the old and the new papers, too. Data was most frequently baseline-corrected relative to 200 ms baseline in new studies (f_{100} = 24%, f_{200} = 52%), and relative to 100 ms in the oldest studies (f_{100} = 44%, f_{200} = 20%).

Unfortunately, descriptions of the order of operations have not become more precise (f_{new} = f_{old} = 64% for papers in which the order of operations could be at least assumed).

Measurement and analysis.

While reporting on the measurement analysis window has changed, the main strategy to choose it has not. The contemporary papers included rationale for choosing analysis window more frequently (f_{old} = 48%, f_{new} = 64% for reports that did have it) and used multiple different arguments to justify the choice more often (f_{old} = 0%, f_{new} = 20%). The main strategy in both groups was visual inspection (f_{old} = f_{new} = 32%). Mean amplitude was the main amplitude measure in both studies (f_{old} = 68%, f_{new} = 64%), while the use of peak amplitude has decreased (f_{old} = 28%, f_{new} = 12%).

Conversely, frequency of reporting on selection of electrodes for the main statistical analysis has not changed (old: $f_{not\ reported}$ = 48%, $f_{inconclusive}$ = 4%; new: $f_{not\ reported}$ = 40%, $f_{inconclusive}$ = 4%), but the most frequently used analysis strategy has. The most common approach of early researchers was to avoid selecting electrodes for analysis by treating all recorded channels as levels of one factor (f_{old} = 28%, f_{new} = 12%), while the contemporary researchers rely on visual inspection more often (f_{old} = 4%, f_{new} = 28%). Like recording montages, analysis montages have also increased (M_{old} = 11.37, M_{new} = 21.76). Consequently, the risk of Type I error has increased with time. This risk was reduced on a different front:

more recent papers had fewer (M)AN(C)OVA models (M_{old} = 10.14, M_{new} = 4.22[18]; papers with only one model: f_{old} = 8%, f_{new} = 32%), as well as fewer ERP components taken from the same waveforms (M_{old} = 2.76, M_{new} = 2.12; papers with only one model: f_{old} = 44%, f_{new} = 80%).

Regarding visualization of spatial distribution, maps have become more widespread (f_{old} = 2, f_{new} = 44%). Topographic distribution analyses have also changed: in the group of older papers, PCA analysis was used in two studies (8%), and it has not been used in the more recent ones. On the other hand, there were four more recent publications (16%) in which LORETA-based analyses were employed.

Overall reproducibility.

Overall, the two groups of studies had similar frequencies of omitting methodological details or presenting them in an ambiguous way. The average contemporary study had some inconclusive information on 1.6 out of 70[19] variables, and some information was omitted in 14.92 cases. Similarly, the older publications had 1.52 variable values with inconclusive and 16 values with missing information.

Providing supplementary methodology materials has become more frequent, although not a norm, in line with the Open Access movement and wider options for storing research data online. Sharing at least brief descriptions of stimuli has become more frequent (f_{old} = 8%, f_{new} = 16%). On top of this, two of the most recent studies (8%) have also published some of their ERP data, albeit only mean component amplitudes in one case.

Summary and implications

In this systematic review, we investigated methodology decisions and clarity of reporting on methodology in peer-reviewed ERP literature. The review encompassed studies published between January 1980 – June 2018 in journals included in two large databases: Web of Science and PubMed, which investigated a well-established component (the N400) in the most commonly assessed population (healthy neurotypical adults), in one of its common modalities (visual images). The review provides insight into study design, data processing, measurement, statistics, visualization of results, and references to supplemental information.

The review aimed to answer six main questions, which will be revisited here.

[18] This difference remains after removing three outliers with more than 40 ANOVAs.

[19] Seventy-four properties were extracted, but publication details, such as paper type (article vs. proceedings) were not included.

How often are descriptions of methods and analyses insufficiently detailed? Which are the principal areas where improvements in reporting practices are necessary?

It would not be difficult to guess which were the most frequently described aspects of the reviewed studies: sample size, number of presented trials, and amplitude measurement window, the type of statistical analyses (e.g. ANOVA) were reported universally or almost universally, with only a few exceptions.

Similarly, *amplitude measure* was reported in 93% of papers, and the *analysis montage* could be extracted from 89% of all papers. These numbers are high, but still concerning, given that these are some of the most important aspects of a study.

At the next level of clarity, there were methodology decisions which were described by the majority of researchers, but there was still a considerable number of papers in which this information was either missing or not adequately described. First, information about the *reference* used for data analysis was provided by about 80% of all researchers. The most frequent issue with reporting on the voltage reference was not providing a description of the recording montage when using the average reference, although, in some cases, details about a mastoid or earlobe reference were omitted. While omitting details about the mastoid reference can be relatively benign, the average reference can differ a lot depending on the recording montage, and it may even be inappropriate to use it depending on the recording montage size and electrode locations. Additionally, in some papers, it was difficult to assess whether the term "linked reference referred to physical linking or averaging. Similarly, *baseline* duration was explicitly described by about 80% of researchers. As described in the section on baseline correction, there were other issues in describing and presenting baseline duration, whose frequencies were not quantified, but which deserve future researchers' attention. *Epoch durations* were provided slightly more often, in about 85% of all cases. It was similar with reporting *impedances* for low input-impedance amplifiers, but descriptions of data quality obtained by high-input impedance amplifiers were provided only in four out of ten papers. *Amplifier manufacturer* and *recording montage* were both provided in about 60% of cases. The latter was often completely left out from the reports, but some of the papers were labeled inconclusive because of conflicting information. Recording montages often have dozens of electrodes, which can make errors easy to overlook, so future researchers may want to make sure to double-check whether all information is correct and consistent. Almost a third of all authors did not describe their methods for eliminating *artifacts* beyond specifying whether they were removed using correction or rejection. Even when more details were given, they were not always sufficient to evaluate and replicate the procedure. Important *decisions about data analysis* – selection of time window(s) and electrode locations for the main statistical analysis – were not justified in about a third of all cases. Moreover, when previous literature was cited as the sole basis for these decisions, in about half of all cases, the cited papers did not support the authors' decisions. In addition, details about the analyses applied to these time windows and electrodes were inconclusive in 4–17% of papers. In some of these cases, parts of information were omitted, but, in others, there was conflicting information between Methods and Results sections. One possible cause of this discrepancy could be the peer review process. Therefore, the future researchers may want to check whether the appropriate changes were made in all parts of the text if a different approach is taken after feedback from reviewers.

Finally, there were aspects of the examined studies which were rarely adequately described, and which warrant urgent attention of researchers and reviewers. In the first place, descriptions of both analog and digital *filters* rarely had information other than cut-off frequency, and it was, too, usually described without specifying what point in the frequency response function it represents. A reconstruction of the *order of pre-processing and measurement steps* could be made in about half of all cases, and in many of these cases, it was only an assumption based on the order in which the operations were described. Finally, we did not quantify this, but it was not possible to determine *how many comparisons* were made in total in some of the studies.

How much variability is there among studies that would be expected to follow similar procedures, because they all investigate the same well-established neurological phenomenon? Which practices are the most prevalent?

There were several points on which most researchers agreed, for better or for worse. The decision which had the most support among researchers was that main effects and interactions are treated as a priori comparisons, and not subjected to any kind of correction (more than 95% of all papers). We demonstrated that this was not appropriate because of the number of comparisons which were made in most experiments. Next, approximately nine out of ten researchers used ANOVA for statistical analyses and avoided using notch filters. About 70% of researchers reported testing for sphericity and applying corrections where necessary, and about 80% of them used the more conservative Greenhouse-Geisser adjustment. Mean amplitude and the mean mastoid reference were used by three quarters of researchers. The latter is especially relevant to future researchers who want to present their data in a way comparable to the previously conducted studies. Finally, analyses based on LORETA were most frequently used to estimate potential sources of ERP components.

The next group of methodological decisions were the ones on which the authors of reviewed publications diverged, but the number of options was not excessive and at least some groups could be identified. Such decisions were baseline (11 different baselines, but 100 ms was used in 43% cases), filter cut-offs (9-18 different cut-offs, but 0.1 and 30 Hz were the most frequent), post hoc comparisons (no correction in 42% of all papers, and 9 different corrections, out of which Bonferroni and Tukey HSD were the most frequent), time window selection strategy (11 strategies, out of which visual inspection was used in about third of all cases), method of selecting electrodes for the main statistical analysis (11 option, none of which was used in more than 14% of all cases), and epoch duration (32 different epochs, but 1000 ms was used in 20% of all cases).

Finally, there were methodological decisions to which almost every team of authors took a different approach. When it comes to artifact correction and rejection, 67 unique pipelines were found, each of them used in only one paper or a handful of publications. However, as long as artifacts are properly eliminated from the trials used for averaging, this variability is not as concerning as the variability found on the average reference locations, analysis montage and the N400 latency. It was not possible to determine how many different electrode montages were used to produce the average reference, because the montages were not described in half of these papers. Still, it can be seen based on the reported montage sizes, that there were at least 14 different montages in 27 papers in which the electrode montage was reported, with as little as 19 or as many as 144 electrodes. Therefore, the topographic distributions of effects obtained from these montages, especially those with fewer than 64 electrodes, differ to an unknown extent. As a result of predominantly data-dependent strategies for the analysis window selection, the researchers measured the N400 amplitude from 69 different latency ranges, 84% of which were used in a single study. Similarly, the N400 effect was determined based on 66 different electrodes combined into 93 unique sets, of 41 different sizes varying between 1 – 144. Furthermore, these sets were subjected to 99 different main analyses. What could a future researcher rely on to make an a priori decision about statistical comparisons, given such variability? We investigated overlap between latencies and electrode locations from different studies and found that the time range in common to most papers was 300–500 ms, and the most frequently analyzed electrodes were F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4. Finally, equally divisive was plotting ERP y-axis, although there are only two options available: 58% of all researchers adopted the 'negative-up" convention, while 40% went for the 'positive-up'. It is telling that after decades of ERP research, the field has not yet reached a consensus on this question.

How often do researchers deviate from guidelines for good practice? Which deviations are the most common?

While "it depends" how many participants and trials are needed for a sufficiently powered study, as Boudewyn et al. (2018) put it, it is safe to say that studies with fewer up to ten participants (11%) and studies in which no more than thirty trials per condition were averaged in some conditions (29%) were underpowered to detect smaller within-group and between-group effects.

Among the recording and pre-processing steps, a few issues were registered. Inappropriately high high-pass and low low-pass filters (either analog or digital) were found in 10% and 8% of papers, respectively. Linked mastoid or earlobe references were used in about a quarter of all studies (assuming that the description of recording with a linked reference was correct), and, in a few studies, ERP data was referenced to the average of all sites despite having as few as 19 electrodes. While all baseline durations were appropriately long (100+ ms), other baseline-related issues were noticed, such as confounding activity in the baseline period. It is difficult to assess prevalence of deviating from the best practices in artifact detection and correction due to limited information available and diversity of methods which were described, but suboptimal strategies were found in some cases (e.g. rejecting trials only on the basis of a fixed base-to-peak threshold).

When it comes to data analysis, deviations from the guidelines for good practice were a norm rather than an exception. When data selection strategies were reported, they were frequently data-dependent, despite relatively stable latency and spatial distribution of the N400. Furthermore, most studies included several analyses of variance, each with multiple factors, and two thirds of studies examined additional time windows and regions in addition to the N400. With this probability of Type I error, it is urgent that ERP researchers reconsider their data analysis strategies. Additionally, there was an overlap between components that were measured and the overt response and/or the next trial in about of 30% of all studies.

Finally, some practices are not deviations from guidelines, but adopting alternatives more broadly may benefit future studies. Two such practices were registered: jittering inter-stimulus interval and boosting their statistical power by lowering impedances even when high-input impedance amplifiers are used.

***

What should be the main takeaway from this study? The goal was not to show that all studies have their issues. It is likely that there are no perfect studies, and ERP data recording, processing and analysis are incredibly complex processes. Moreover, these very standards we have today, which were cited in this study, result from continuous endeavors by the ERP research community to improve methods and analyses of ERP data. Many problems presented here not unique to ERP research – on contrary, they are shared with similar fields of study, such as fMRI, psychophysiological recordings, and, in some respects, even behavioral research. This study, therefore, serves to highlight some common issues, to provide guidance for a priori time window and electrode selection, and to advocate for more rigorous methodology and more comprehensive reporting in future.

This systematic review, although extensive, is far from exhaustive. Picture-evoked N400 is not the only ERP measure, and many methodology decisions were not considered in this review – from statistical power, to study design and hypotheses, participant exclusion criteria, compliance of graphs with recommendations for appropriate visualization of ERP data, details of more complex statistical analyses, and others. These questions remain to be explored in future studies.

03 Varying processing and analysis parameters

Is the observed lack of coherency and transparency really an issue?

Hypotheses

We expected that the experimental results (Analysis 1) and statistical power (Analysis 3) would depend on the processing and analysis pipeline, even though individual steps were varied within limits found in the existing picture-evoked N400 literature. In addition, we expected statistically significant interactions between processing decisions and the experimental factor, i.e., we expected processing steps to influence N400 effect size (Analysis 2).

If predicted effects of a processing/analysis decision are registered, it would show that this decision can influence outcomes of a study, and that researchers and reviewers should be cautious when considering these decisions. On the other hand, given that only one dataset was analyzed, with its specific effect size and noise, we did not test for equivalence between processing pipelines, as evidence of similarity between outcomes in this dataset would not mean that the same decisions would not have more substantial consequences under different circumstances.

Method

Description of the study which will be used for analyses

The study chosen for our analyses comes from the field of linguistic relativity research, and it was conducted by Boutonnet, McClain and Thierry (2014). It has been chosen for several reasons. The experiment features a simple, classical N400 priming design, similar to most other papers. Second, the EEG signal is of high quality and it was recorded with minimal analog processing. Third, offline processing conducted by the authors was in line with a typical pipeline and within limits recommended by the guidelines. Namely, most processing steps overlap with the one of the two most common practices registered in the systematic review. Exceptions are analysis window and electrode choice. Additionally, there are two practical advantages: the equipment used in the experiment is identical to the one that our laboratory has, and the authors were willing to provide additional information, not provided in the article. The study will be briefly described in the paragraph that follows, and, afterwards, more detail will be given about its aspects relevant for this study.

The question of interest for Boutonnet et al. (2014) was whether arbitrary, language-specific relations between words can impact non-verbal memory representations. One example of language-based relations are word compounds that tie two words that are not semantically, functionally or in any other way related, such as seahorse or butterfly. To study the effects of these artificial conceptual relationships, the authors employed a non-linguistic picture relatedness task featuring picture pairs of compound elements presented in both original (e.g., sea – horse) and reverse order (e.g., horse – sea), in addition to semantically related (e.g., butter – cake) and completely unrelated (e.g., butter – boy) picture pairs.[20] During the task performance, ERPs were recorded from the participants, and N400 amplitudes were examined. The results revealed that showing compound pairs presented in the reverse order reduced N400 amplitudes to a statistically significant degree. The authors concluded that the results demonstrate that semantic memory associations can be formed and influenced by language and its idiosyncratic relations.

Participants.

There were 16 participants (nine female, seven male, 21.9 ± 0.9 years old), native speakers of English and students at the Bangor University School of Psychology. They received course credits for their participation. The local ethics committee reviewed and approved the study.

Stimuli.

Stimuli were based on 50 compound words, each consisted of two highly imageable and familiar constituents (e.g., butterfly). One hundred photographs showing prototypical representations of each compound constituent were selected. Further information about stimuli properties and their full list can be found in the article by Boutonnet et al. (2014).

Procedure.

At the beginning of their session, participants received instructions and signed a consent form. The testing was conducted individually in a quiet room.

The pictures were arranged into pairs to create a total of 200 trials, divided into four fully rotated experimental conditions, each consisting of 50 trials. Each participant saw all trials twice, resulting in a total of 400 trials, or 100 trials per condition. The conditions were: semantically related (Related), related by a compound (Compound), related by a compound but presented in reverse order (Reverse), and Unrelated. The Compound and Reverse conditions both featured compound constituents paired together, but they were presented in the reversed order for the latter condition. The Related and Unrelated conditions were created by rearranging pairs – assigning to each compound's beginning another compound's ending. In the Related condition, the resulting pairs were semantically related, and in the Unrelated condition, there was no relationship between the first and the second picture in any pair.

Each trial started with a fixation cross lasting 250 ms, succeeded by the first picture of a pair, presented for 500 ms. A random variable inter-stimulus interval followed, lasting 400, 450, 500, 550 or 600 ms, averaging to 500 ms. The second picture appeared after the inter-stimulus interval and remained for a maximum of 3000 ms or until participant response.

[20] The last two situations (semantically related and completely unrelated pairs) represent the standard N400 design, used to produce the basic and thoroughly replicated N400 effect.

Trials were separated by a blank screen lasting 500 ms on average. Block order was counterbalanced, and stimulus presentation was randomized.

The participants were instructed to respond whether each pair was semantically related or not. Only the pairs from the Related condition counted as related, and the other three conditions were considered unrelated. Participants were not informed about the presence of compound pairs.

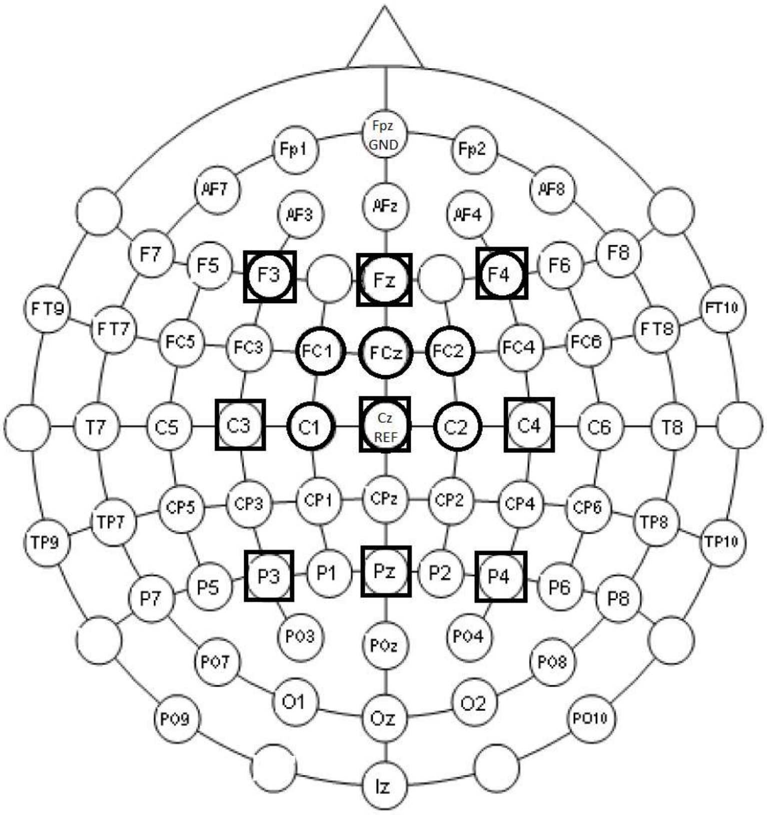


Figure 3.1. Electrode recording and analyses montages. Electrodes that were used in statistical analysis by Boutonnet et al. (2014) are marked with bolded circles, while bolded squares show an alternative montage that will be used to compare results. Recording reference was Cz, which was later changed to the average of all sites.

Electrophysiological recording.

The EEG was continuously recorded using Compumedics Neuroscan 64-channel SynAmps amplifier (https://compumedicsneuroscan.com), and 64 Ag/AgCl electrodes, placed in a cap in locations defined by the extended 10–20 convention (American Electroencephalographic Society, 1991). Fpz electrode served as ground, and Cz as reference (see Figure 3.1 for exact locations). Additional electrodes were attached on the outer sides of both eyes, as well as above and below the left eye, in order to monitor eye movements and blinks. Impedances were maintained below 5 kΩ for the scalp electrodes and below 10 kΩ for the EOG electrodes. Data was recorded at a rate of 1 kHz and filtered online with a broad bandpass filter of 0.05-200 Hz (-6 dB, 12 dB/octave slope) to avoid amplifier saturation and aliasing.

EEG offline data processing and measures.

Offline processing was conducted in Scan 4.4 (Compumedics Ltd. https://compumedicsneuroscan.com). Data were filtered offline using a bandpass filter, with a high-pass half-amplitude cut-off at 0.1 Hz (12 dB/octave slope) and a low-pass half-amplitude cut-off at 30 Hz (48 dB/octave roll-off). Eye blinks were then corrected using the algorithm provided by Scan 4.4, which is based on a regression approach developed by Gratton, Coles, & Donchin (1983). Epochs of -100 to 1000 ms were created next, followed by baseline correction using a 100 ms pre-stimulus period. Individual epochs were averaged, and then referenced offline to the average of all scalp electrodes.

N400 measurement and data analysis.

The N400 component was operationalized as mean amplitude using the 350-480 ms window, which was selected based on variations in the global field power (GFP). Nine electrodes in a 3x3 montage were chosen for analysis: F3, Fz, F4, FC1, FCz, FC2, C1, Cz, C2 (see Figure 1). The montage was determined a priori, based on the expectation that N400 would be maximal over central electrodes. It was not further specified how GFP was used to select the latency range or why this specific combination of electrode sites were chosen to represent the theoretical region of interest. Other components (P1, N1) and peak latencies were also mentioned, but they were not analyzed further.

A three-way, 4x3x3 repeated-measures ANOVA was employed to examine N400 effects. Mean amplitudes were used as the dependent variable and the factors were: condition (Related, Unrelated, Reversed and Compound), anteriority (anterior, central, posterior), and laterality (left, center, right). The four experimental conditions were compared with one another using post-hoc paired t tests on amplitude measures based on linear derivations of all chosen electrode sites. Behavioral responses were also examined.

Results.

ANOVA results showed that all of the main effects were significant: condition [F(3,45)=6.84 ,$p<0.001$, $\eta^2_p=0.31$], anteriority [F(2,30)=23.6, $p<0.001$, $\eta^2_p=0.61$] and laterality [F(2,30)=8.91, $p<0.001$, η^2_p =0.37]. There were no significant interactions. Post-hoc tests revealed significant differences between Unrelated and two other conditions: Related [t(15)=−3.04, p<0. 05] and Reversed [t(15)=2.3, $p<0.05$], while Unrelated and Compound conditions did not differ significantly. On the other hand, the Compound condition differed significantly from the Related condition [t(15)=−3.6, $p < 0.05$].

Our processing and analyses

Offline data pre-processing was conducted in Scan 4.4 (Compumedics Ltd. https://compumedicsneuroscan.com). Measurement and Monte Carlo simulations were performed in MATLAB (Mathworks, Inc. https://www.mathworks.com). Analyses 1 and 2 were performed in JASP 0.11.1.0 (JASP team, 2019). Data was visualized using MS Office 365 Excel, version 1910 (Microsoft Corp., www.microsoft.com).

Raw data from the study by Boutonnet et al. (2014) was processed and analyzed using different values for each processing step, to examine the most common options and variability found using the systematic review approach. There are 864 possible pre-processing and analysis pipelines only based on the methodological choices which were selected for investigation. For this reason, each decision was varied independently, while other parameters were held constant at the settings chosen by Boutonnet et al (2014). As a result, 14 different pre-processing and analysis pipelines were performed, and the pipeline identical to the one described by Boutonnet et al. (2014) served as a standard for comparison. They are summarized in Figure 3.2. and described in more detail below.

ERP data pre-processing.

The first step was **filtering**. High-pass and low-pass cut-off points were varied, while other filter settings were the same as those used by Boutonnet et al. (2014). Zero-phase FIR filters with 12 dB/octave slope for high-pass and 48 dB/octave for low-pass was implemented. Please note that all cut-off values in this study denote half-amplitude points. As seen Chapter 2, offline high-pass filter cut-off varied between 0.01 and 1 Hz in the existing literature, with 0.1 Hz being the most common cut-off point.[21] Boutonnet et al. (2014) also used 0.1 Hz. Therefore, digital high-pass filter was: (1) not implemented in one processing pipeline, leaving the initial, online, 0.05 Hz half-amplitude cut-off, (2) implemented with a 0.1 Hz cut-off, and (3) implemented with a 1 Hz cut-off. Low-pass filter cut-offs varied between 5.5 Hz and 100 Hz in the picture-related N400 literature, and the most common choice was 30 Hz. This value was used in the study by Boutonnet et al. (2014). Consequently, we applied low-pass filters with 5.5 Hz, 30 Hz and 100 Hz. Note that both 1 Hz high-pass and 5.5 Hz low-pass cut-off are outside of the range optimal for cognitive ERP components such as the N400 (Luck, 2014), but these were values that were found in published papers. In case of high-pass filters, it has been shown that half-amplitude cut-offs equal to or larger than 0.3 Hz can reduce the amplitude of the N400 component and induce artifactual effects in the P2 and P600 latency ranges (Tanner et al., 2015).

After filtering, **artifact correction** was implemented, except in the pipeline in which artifact rejection was used. Given that, if conducted appropriately, all artifact correction methods should give comparable results, artifact correction method used by Boutonnet et al. (2014) – the method developed by Gratton et al. (1983) – was contrasted with artifact rejection. **Artifact rejection** was applied after extracting epochs and performing baseline correction. It was based on manually set base-to-peak thresholds, and two researchers who jointly set the thresholds were blind to conditions. The researchers reviewed automated

[21] Please note that most papers did not specify whether offline filter cut-offs were half-amplitude or half-power. They also typically did not include information on slope. Therefore, it was not possible to convert all values into the same format.

threshold-based decisions and manually made corrections to keep or discard additional trials. Blinks and other large artifacts were eliminated this way. This choice was not the most frequently registered in the systematic review. The variability of artifact rejection procedures was large, and there was not one solution that could be singled out as more frequent than others. Instead, we chose an option which is both recommended in methodology literature (Luck, 2014) and available in our processing software package.

The third step was segmenting data into **epochs** of -350–1000 ms. Prestimulus period was chosen to accommodate all baseline lengths, and poststimulus period was the same as the one chosen by Boutonnet et al. (2014). The only exception was the pipeline in which artifact correction was replaced with rejection. In this pipeline, the epoch was shortened to -100–700 ms. Namely, with a -350–1000 ms epoch most trials would be discarded in the trial rejection pipeline, and the shorter epoch included both the baseline and analysis window used in the pipeline in which artifact rejection was used.

After segmenting, **baseline correction** was performed, relative to 100 ms, 200 ms, and 350 ms prestimulus period. We excluded 200 ms and 350 ms durations if they encompassed confounding ERP activity, as this would lead to artifactual differences between ERP waveforms (Luck, 2014). Prestimulus baseline durations in the reviewed N400 papers varied between 100 to 350 ms. Duration of 100 ms was not only the low extreme, but also the most frequent choice and the choice used by Boutonnet et al. (2014). Given that these three values were all equal, we added a 200 ms prestimulus baseline, which was the second most common option, to include a middle ground length.

After baseline correction, **individual averages** were created for all experimental factor conditions. In the pipelines in which eye blink correction was performed, all trials were averaged. In case of three participants, 1, 5 and 56 trials were not recorded, presumably due to a technical issue. For this reason, an average of 99.28 trials were averaged per condition when artifact correction was performed. After artifact rejection, about a quarter of all trials were eliminated, mostly due to eye blinks, leaving an average of 74.82 trials per condition. The number of trials did not differ considerably between conditions. More details about number of trials averaged together is provided in Table 3.1.

Table 3.1. Descriptive statistics of the number of trials averaged together in pipelines in which artifact correction and artifact rejection were performed.

Artifact elimination method	**Condition**	**Number of averaged trials per condition**			
		M	**SD**	**Min**	**Max**
correction	Unrelated	99.12	3.14	87	100
	Compound	99	3.14	87	100
	Reverse	99.88	3.87	84	100
	Related	99.12	3.39	86	100
rejection	Unrelated	72.81	21.40	32	98
	Compound	75.12	18.85	42	95
	Reverse	74.94	16.62	44	95
	Related	76.43	16.78	38	96

Finally, all data was **re-referenced** either to the average of all sites, like in the analysis by Boutonnet et al. (2014), or to the average of the mastoids. Results of the systematic review showed that average mastoids were the most common reference, followed by the average reference.

Additionally, for the purpose of Monte Carlo analyses (see Analysis 3), measurement and re-referencing took place before averaging, because subsets of trials were averaged together in each iteration. Notably, this change in the order of steps did not have any effect on the results, because all of these operations are linear. The only exception is peak amplitude[22], which was excluded from the Monte Carlo analysis for this reason.

N400 measurement.

The N400 **amplitude** was measured using *peak and mean* values within *four windows*: the latency range used by Boutonnet et al. (2014), 350–480 ms, the range most frequently used in the picture-related N400 literature, 300–500 ms, and the shortest and longest windows found in literature, centered on the same time point as the window used by Boutonnet et al. (2014): 398–424 (24 ms), and 185–635 (450 ms)[23].

Two **analysis montages** of nine electrode sites were used. Both are shown in Figure 3.1. The first is montage used by Boutonnet et al. (2014) (F3, Fz, F4, FC1, FCz, FC2, C1, Cz, C2). The second is comprised of the electrode sites most commonly analyzed across all studies, as found using systematic review approach: F3, Fz, F4, C3, Cz, C4, P3, Pz, P4.

Variables.

Besides condition – the main experimental variable, steps of data processing and statistical analysis were treated as factors in this study.

The following factors were included: offline filter high-pass cut-off (levels: no digital filter, 0.1 Hz, and 1 Hz half-amplitude), offline low-pass filter cut-off (levels: 5.5 Hz, 30 Hz, 100 Hz), eye artifact elimination method (levels: correction, rejection), baseline correction time window (levels: 100 ms, 200 ms, 350 ms, with longer durations potentially excluded), N400 amplitude measure (levels: mean, peak), N400 measurement window (350–480, 300–500, 398–424, 185–635), reference (levels: average mastoids, average of all scalp sites), analysis montage (levels: montage used by Boutonnet et al. (2014), the nine most frequent sites found using systematic review approach).

[22] Because finding a peak of a component is not a linear operation, measuring a peak separately on each trial, and then averaging peaks, and measuring a peak from averaged trials would not produce the same result.

[23] The longest window recorded in the systematic review of the N400 literature was 600 ms (200-800 ms), but this window was analysed separated into six 100 ms latency ranges. The chosen analysis window duration of 450 ms was the longest range that was analysed as an integral unit.

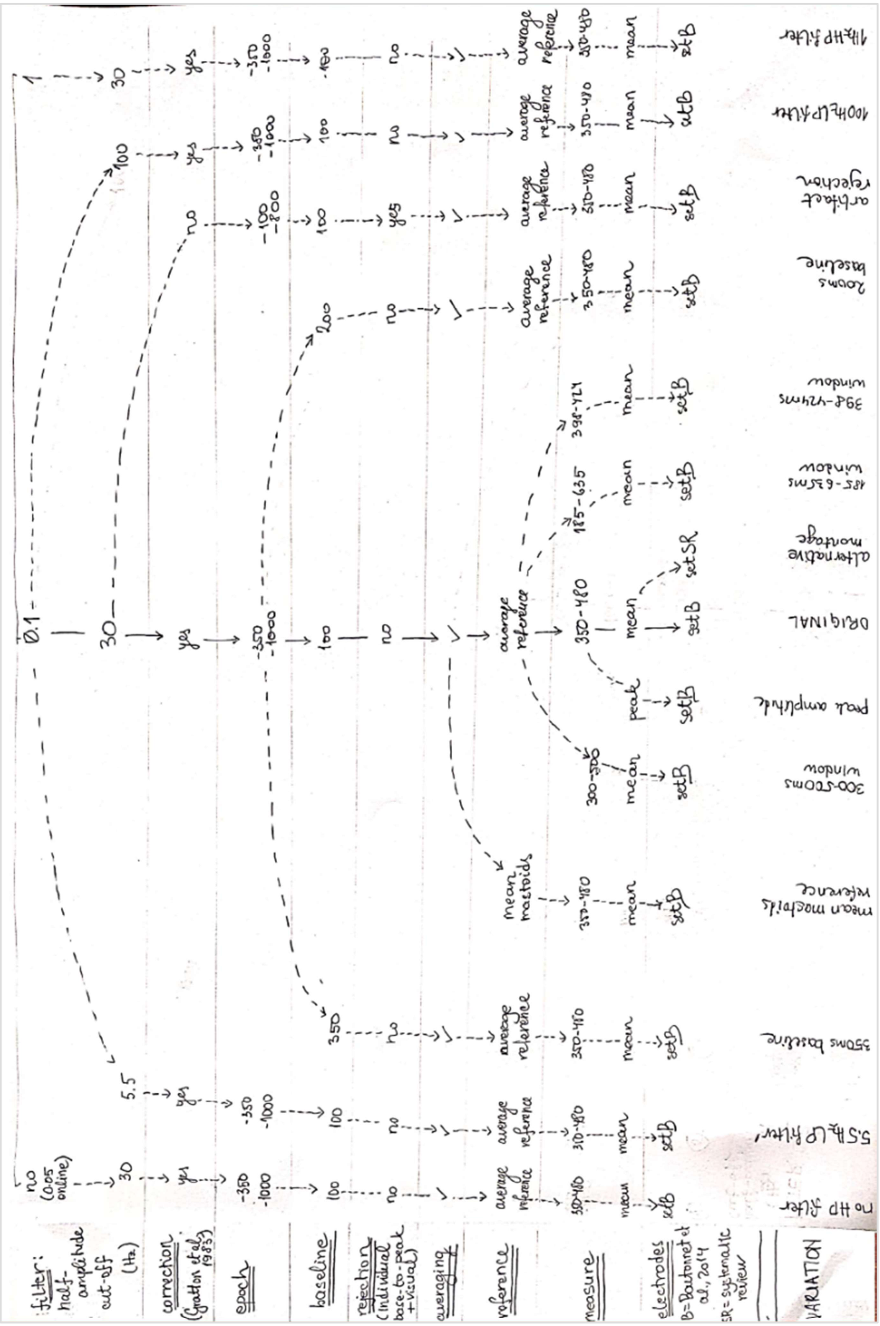


Figure 3.2. All pre-processing and analysis pipelines. The main pipeline, which was the same as the pipeline used by Boutonnet et al. (2014), is in the middle. All other pipelines are made by varying one step from the main pipeline. These variations are summarized in the bottom of the Figure. For purposes of Monte Carlo simulations, the order of steps in these pipelines was changed so that averaging trials was the last step in the pipeline, and this was performed for all pipelines except for the pipeline in which the amplitude was measured as a peak voltage within a time window.

Laterality and anteriority factors were not included. Instead, amplitudes registered on the selected electrode sites were averaged together in all analyses, for two reasons. First, we are interested whether the main N400 effect would be replicated with all analysis pipelines. To limit the number of comparisons, we did not plan to investigate main and simple effects or interactions involving location factors (see Data analysis), and any analyses of other factors would yield the same results as if we had used two-way ANOVAs on the average of all nine locations. Second, averaging all sites together allowed us to treat analysis montage as one factor with two levels, and consequently to directly compare recordings with different analysis montages.

In addition, number of trials per condition was varied in Monte Carlo simulations.

Data analysis

Analysis 1: Changes in significance.

In the most basic test, the results of new analyses were compared to the original study's findings. The goal was to examine if the variations in analysis pipeline could produce conclusions about the experimental effects which are different than the original ones.

One-way ANOVAs with condition as the independent variable and mean amplitude as the dependent variable was conducted on each dataset resulting from the new analyses. Statistical significance of the main effect of condition from each analysis was compared to the findings of Boutonnet et al. (2014), while post hoc comparisons were not examined, and effects of electrode site were not be calculated, as outlined in the previous section.

In addition to the p-value, experimental factor effect sizes (η^2) were compared.

Analysis 2: Interactions between condition and methodological decisions.

Furthermore, eight two-way ANOVAs were conducted,[24] with each data processing parameter as a factor in addition to the experimental condition (relatedness). In each of these ANOVAs, the average amplitude recorded on all nine electrodes used by Boutonnet et al. (2014) was the dependent variable. The only exception was comparison of electrode sites, in which electrode sites were varied, so the two-way ANOVA included two factors: experimental condition and analysis montage, with two levels: average amplitude on the nine sites used by Boutonnet et al. (2014) and average amplitude on the nine sites chosen for comparison.

[24] Conducting an omnibus ANOVA including all factors would allow comparing interactions between processing steps as well. However, this would require performing all 864 possible combinations of processing and analysis steps. This would not be practically feasible, and it would result in a tremendous increase in the number of potential comparisons, and therefore require lowering the threshold of statistical significance and, consequently, statistical power to detect main factors of interest.

To avoid inflating Type I error probability, only two-way interactions between (1) individual data processing parameters and (2) the experimental condition were considered. Because there were eight planned comparisons, the overall alpha level $\alpha = 0.05$ was adjusted using Šídák method to control for Type I error rate increase, resulting in $\alpha_{\text{corrected}} = 0.00639$. These comparisons were chosen because they reflect effects of processing parameters on the main experimental effect. Exploring main effects of processing parameters on the N400 component would also be relevant in case of the variables which should not influence the component value significantly (e.g. filter cut-off value). However, these analyses would have almost the same meaning as two-way parameter × condition analyses, and two way-analyses are closer to the subject of interest in the study. This is due to the fact that two-way interactions reflect the effect of the parameters on the N400 *effect***,** while main effects reflect the effect of parameters on the N400 *component***.**

Anaylsis 3: Monte Carlo simulations.

The third approach included conducting Monte Carlo simulations to determine how much each of the parameters influences statistical power to detect experimental effects, as well as how the number of trials affects differences between data processing set ups. This approach has been used in ERPology to determine effects of electrode impedances (Kappenman & Luck, 2010) and high-pass filter settings (Tanner et al., 2015) on ERP experimental results.

Monte Carlo analyses involve simulating a large number of experiments with varying numbers of trials or participants by taking random subsets of data from a study. As a result, it is possible to get an estimate of statistical power by calculating probabilities of achieving statistical significance for each subsample size. In our study, we simulated experiments with varying numbers of trials, and for each of them we performed omnibus ANOVAs resulting from different processing pipelines. To obtain a robust estimate of the probability of significant effects ($p<0.05$), 1000 experiments were simulated for each of the following subsample sizes: 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70 trials. The only exception was the pipeline with trial rejection, which did not have enough trials for simulations with subsets larger than 30 trials.

Given that the experimental effect includes a long-established difference between semantically matching and unrelated pictures, we expected it to be significant. In that case, the probability of reaching statistical significance provides information about the signal to noise ratio, and, consequently, a platform to compare different processing strategies.

As noted earlier, peak amplitudes were not subjected to Monte Carlo simulations.

***

In all analyses of variance described above, we corrected degrees of freedom using Greenhouse-Geisser method when sphericity was violated, as indicated by Mauchly's test. Because this method is overly conservative, we additionally consulted the results obtained using Huynh-Feldt method, and when the two methods produced opposite conclusions about statistical significance, we conducted ANOVA with degrees of freedom adjusted using the average of the two ε correction factors (Stevens, 2009). In each case, the conclusions based on the average ε correction factor were the same as when Greenhouse-Geisser adjustment was used.

The only exception were Monte Carlo simulations, in which all analyses were performed with Greenhouse-Geisser correction, due to practical considerations.

Pre-registration and discrepancies between pre-registration and the analyses that were conducted

Expectations and method for this study were pre-registered on the Open Science Framework platform prior to examining and processing data which was used for the analyses. The project page for the study, which includes the pre-registration document can be accessed via the following link: https://osf.io/6nqxy/ . All pre-registered decisions on data processing and analyses were made before examining data and while blind to experimental conditions. This ensure this, codes for conditions were obtained from the authors of the study after finishing the analysis plan.

However, after examining data, we found a few, mostly minor, discrepancies between the dataset we used and the publication in which the study was described (2014). Due to these discrepancies, it was necessary to make some adjustments to the pre-registered pre-processing and analyses pipeline. Discrepancies we found, changes to the pre-registered pipeline, and justifications for these changes will be described in the end of the Methods section.

First, Boutonnet et al. (2014) state in their Methods section that 51 stimuli per condition were used, and we based our pre-registration on this information. However, the stimuli list provided as a supplement to the publication, shows that there were 50 stimuli per condition. The raw EEG recordings from the study had 100 stimuli per conditions, indicating that the participants viewed all stimuli twice, a practice not uncommon in ERP research. We used the fact that there were more trials than reported in the publication to expand the Monte Carlo simulations from initially registered 10-40 trials to cover 10-70 trials range. Additionally, we decreased the step between simulated sample sizes from 10 to 5 to create more fine-grained curves.

Second, in the paper, the researchers describe that eye blinks usually only occurred after response due to an unspecified instruction given to the participants, and they do not report performing artifact rejection. Based on this information, we did not expect many trials to be rejected using the procedure we pre-registered. In the pre-registration document, we reported a plan to perform artifact rejection on epochs of the same duration as in other pipelines, based on a threshold set individually for each participant. However, inspection of uncorrected epochs showed that participants regularly blinked after stimulus presentation. As a result, the epoch which we pre-registered included blinks often, sometimes on most trials. Due to this issue, most trials would be eliminated by artifact rejection strategy which we pre-registered. In order to overcome this issue, we shortened the epoch to -100–700 ms duration, and visually inspected epochs to improve the results of automatic artifact correction. Note that in the pre-registration, we based our decision on the ERP methods book by Steven Luck (2014), and that manual revision of unsatisfactory automated rejection decisions is recommended by the same source as long as the researchers performing this task are blind to conditions.

Due to an oversight on our part, we did not pre-register which method for correcting data for sphericity violation will be used.

Results

Waveforms and descriptive statistics

ERP waveforms obtained from each pre-processing and analysis pipeline are provided in Graph 3.1, while means and standard deviations of the N400 component amplitudes are shown in Table 3.2.[25]

Compared to the analysis pipeline described by Boutonnet et al. (2014), some variations produced an increase or decrease in the magnitude of differences[26] between conditions, while others had effects on between-subjects variability. We will briefly describe the effects on variations in the pipeline on the magnitude of the difference between Related and Unrelated conditions, which is the standard, and thus reliable, N400 effect, as well standard deviations of amplitudes measured in all four conditions.

The magnitude of the difference between Related and Unrelated conditions, but also standard deviations in all conditions were reduced when high-pass filter cut-off increased to 1 Hz. Graph 3.1.b shows how the difference between the Related and Unrelated conditions was reduced in this pipeline version, among other changes. The same effect was observed when the set of electrodes chosen based on the systematic review was used instead of electrodes selected for analysis by Boutonnet et al. (2014) (Graph 3.1.e). This change primarily resulted from the polarity inversion of all ERPs in the posterior part of the scalp, compared to the frontal part (see Appendix J). Because the largest difference was observed starting around 350 ms (Graph 3.1.s), broadening the measurement window from 350–480 ms to 300–500 ms and 185–635 ms had the same result, although to a smaller extent compared to the above-mentioned variations.

Conversely, both the magnitude of the difference between conditions and the variabilities were larger when low-pass filters increased from 30 Hz to 100 Hz (Graph 3.1.d),

[25] Additionally, Appendix J shows ERP waveforms obtained from all scalp locations from two pipelines: the one which described by Boutonnet et al. (2014), and the one in which data were re-referenced to the average mastoid reference. One noteworthy aspect which can be seen in these waveforms is that all waveforms had a similar pattern, which was also to an extent present in the HEOG electrode, indicating possible presence of residual noise. We observed the same pattern in all pipelines, as well as in grand averages provided to us by study authors.

[26] In line with terminology used by Boudewyn et al. (2018), the term *effect magnitude* is used here to refer to the absolute differences in voltage, as opposed to *effect size*, which is used when differences are viewed relative to the amount of variability.

when the shortest measurement window (398–424 ms) was used (Graph 3.1.s), and, especially so, when voltages were re-referenced to mean mastoid reference (Graph 3.1.f).

The amplitudes measured after applying only online filters had similar amplitude averages and standard deviations to the ones obtained when the procedure outlined by Boutonnet et al. (2014) was followed. Similarly, lowering low-pass filter to 5.5 Hz affected neither the difference in means, nor standard deviations, but it had an obvious effect on the overall waveform, as it can be seen in the Graph 3.1.c.

Increasing baseline duration (Graph 3.1.g, 3.1.h) decreased amplitude variabilities, but when baseline was extended to 350 ms, the difference between conditions also decreased, due to activity during the early prestimulus period which affected the positioning of ERP waveforms on the y-axis.

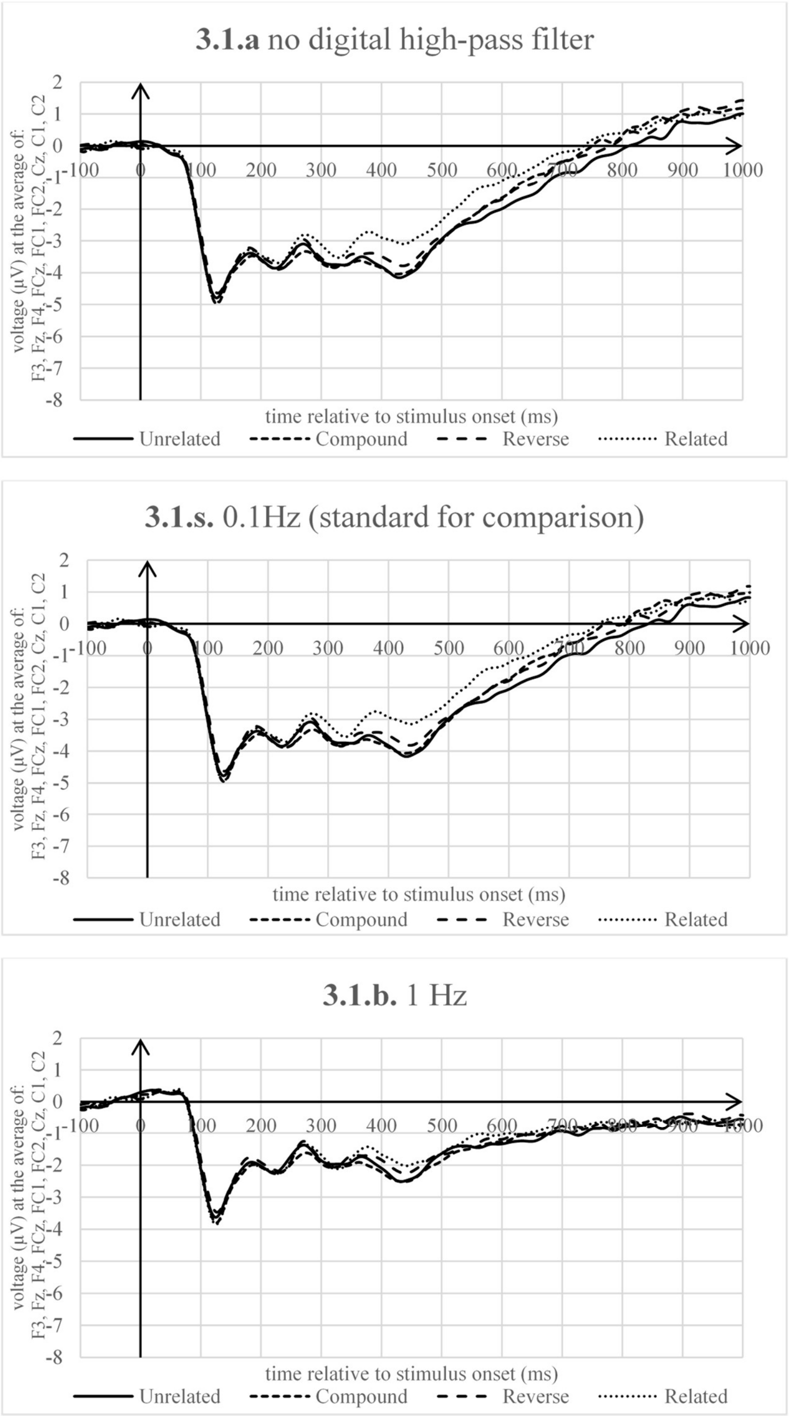
Table 3.2. N400 amplitude descriptive statistics for all pipelines.

Analysis pipeline variations	**Unrelated**		**Compound**		**Reverse**		**Related**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
standard	-3.81	2.63	-3.82	2.88	-3.55	2.45	-2.96	2.74
high-pass filter								
no digital filter	-3.80	2.62	-3.80	2.89	-3.52	2.47	-2.91	2.77
1.0 Hz	-2.14	1.44	-2.24	1.49	-1.94	1.29	-1.76	1.42
low-pass filter								
5.5 Hz	-3.91	2.64	-3.91	2.87	-3.64	2.47	-3.10	2.75
100 Hz	-3.82	2.72	-3.81	2.97	-3.55	2.53	-2.30	2.83
analysis window								
300–500 ms	-3.73	2.54	-3.75	2.80	-3.53	2.40	-3.03	2.69
185–635 ms	-3.24	2.21	-3.21	2.47	-3.08	2.02	-2.65	2.29
398–424 ms	-3.97	2.76	-3.98	3.03	-3.63	2.61	-3.00	2.99
electrodes								
alternative set	-0.74	1.37	-0.75	1.43	-0.60	1.24	-0.32	1.43
reference								
mastoid	-4.98	4.57	-4.94	5.05	-4.31	4.34	-3.13	5.20
baseline duration								
200 ms	-3.64	2.48	-3.59	2.76	-3.43	2.39	-2.78	2.62
350 ms	-3.23	2.14	-3.18	2.43	-3.15	2.03	-2.45	2.30
artifact approach								
rejection	-3.84	2.83	-3.74	2.81	-3.46	2.55	-2.82	2.03
amplitude measure								
peak	-4.75	2.66	-4.59	2.99	-4.29	2.43	-3.75	2.65

Note: M = mean; SD = standard deviation, with a normalization factor N; standard = pre-processing and analysis the same as the ones described by Boutonnet et al. (2014).

3.1. Grand average ERP waveforms

Digital high-pass filter half-amplitude cut-off



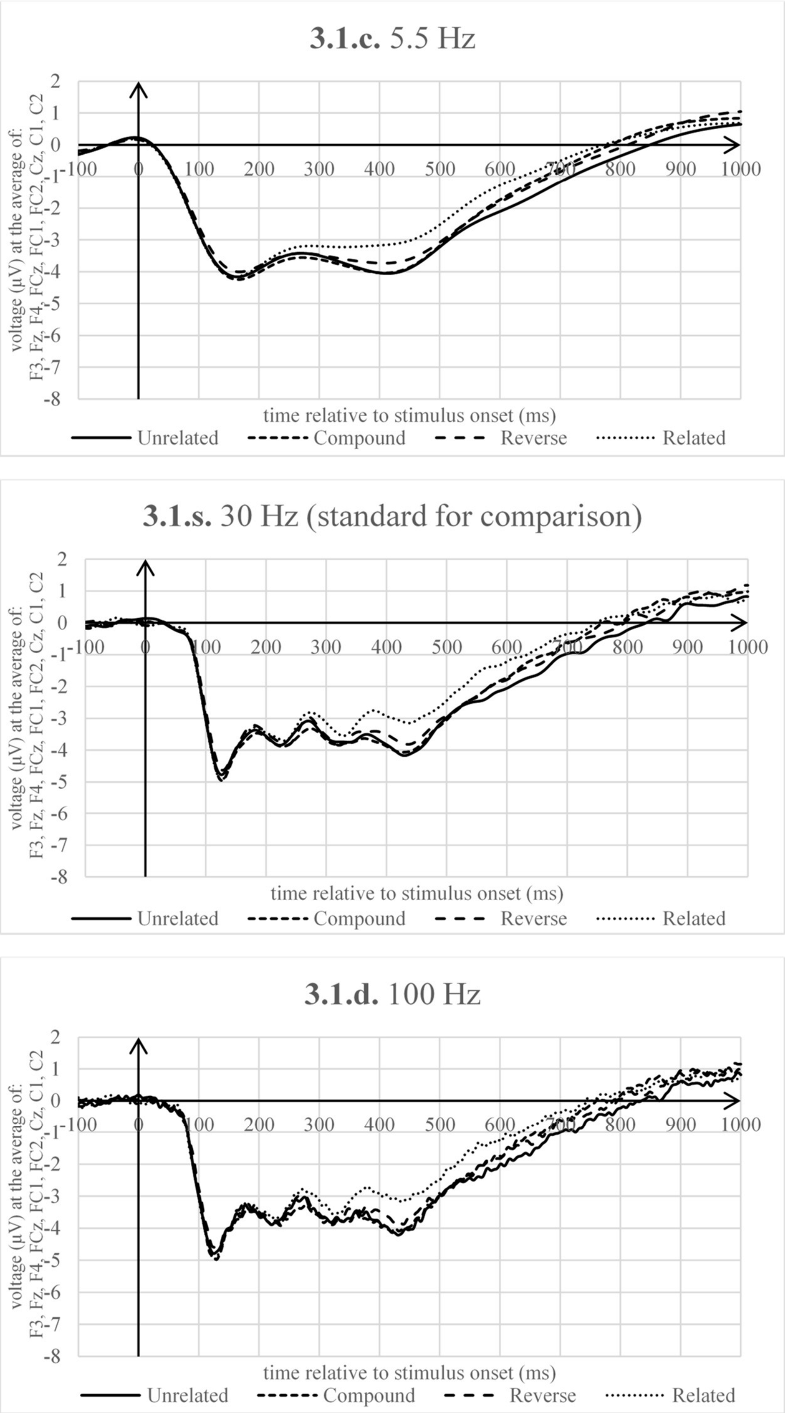
3.1.a no digital high-pass filter

3.1.s. 0.1Hz (standard for comparison)

3.1.b. 1 Hz

3.1. Grand average ERP waveforms

Digital low-pass filter half-amplitude cut-off

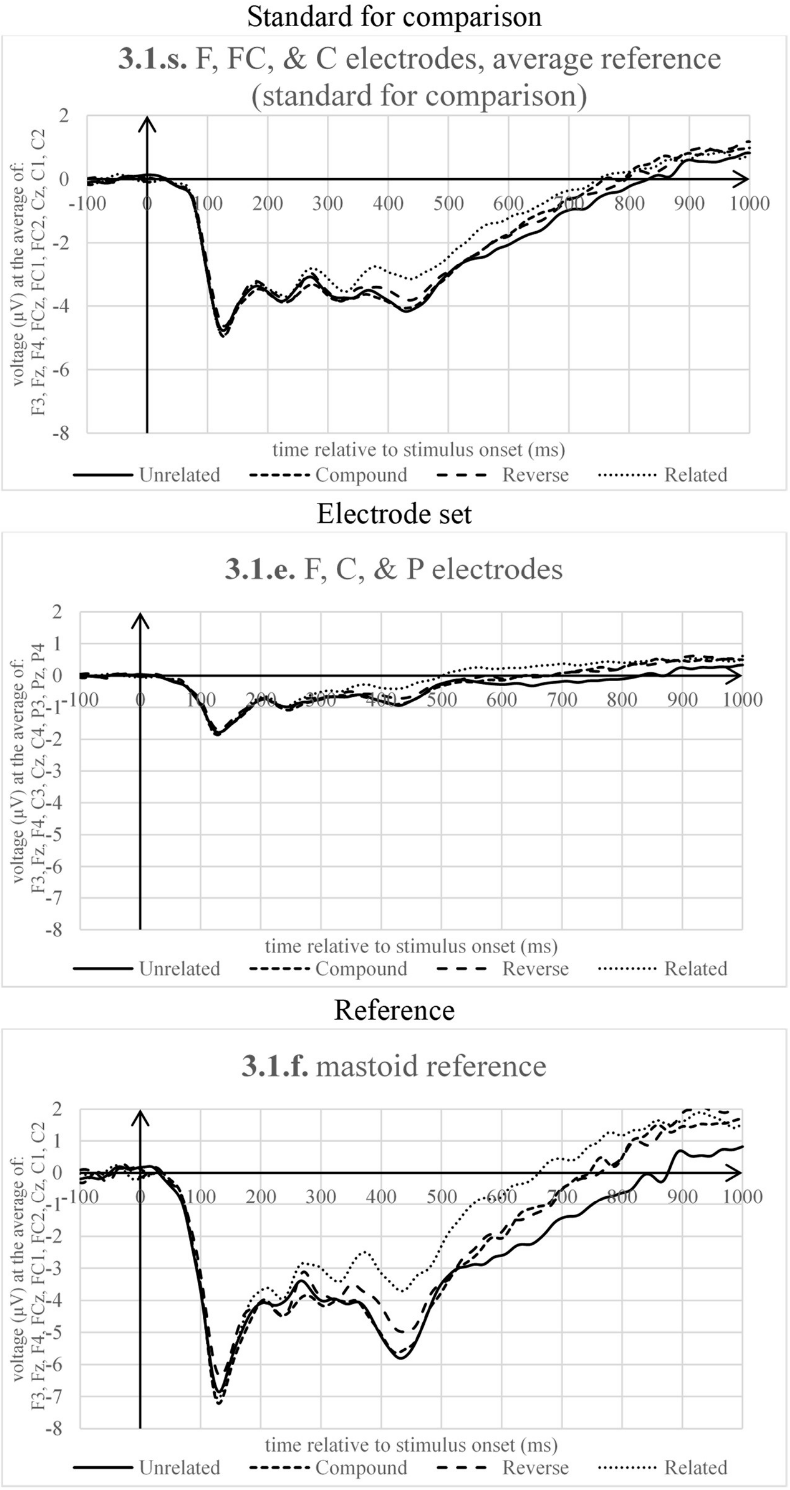


3.1.c. 5.5 Hz

3.1.s. 30 Hz (standard for comparison)

3.1.d. 100 Hz

3.1. Grand average ERP waveforms



Standard for comparison

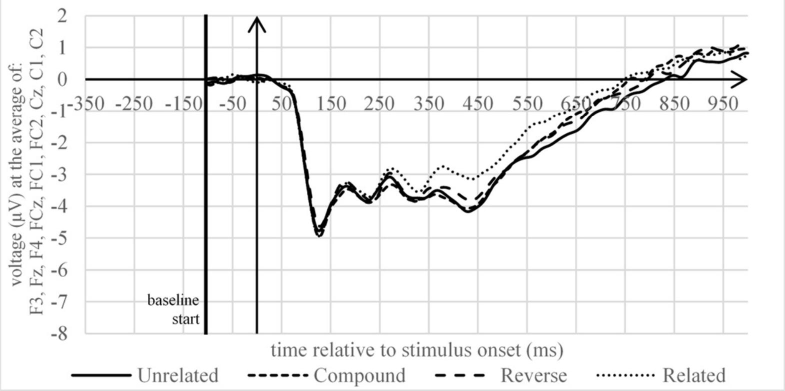
Electrode set

Reference

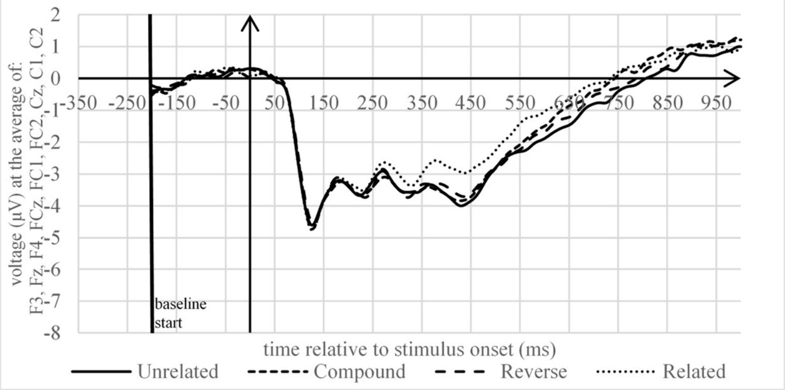
3.1. Grand average ERP waveforms

Baseline duration

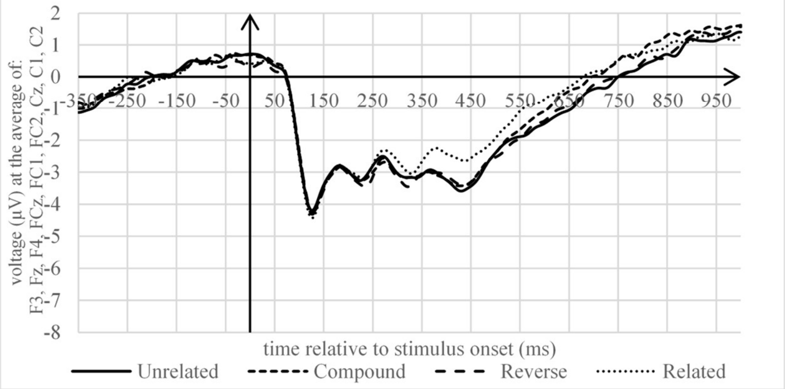
3.1.s. 100 ms (standard for comparison)



3.1.g. 200 ms

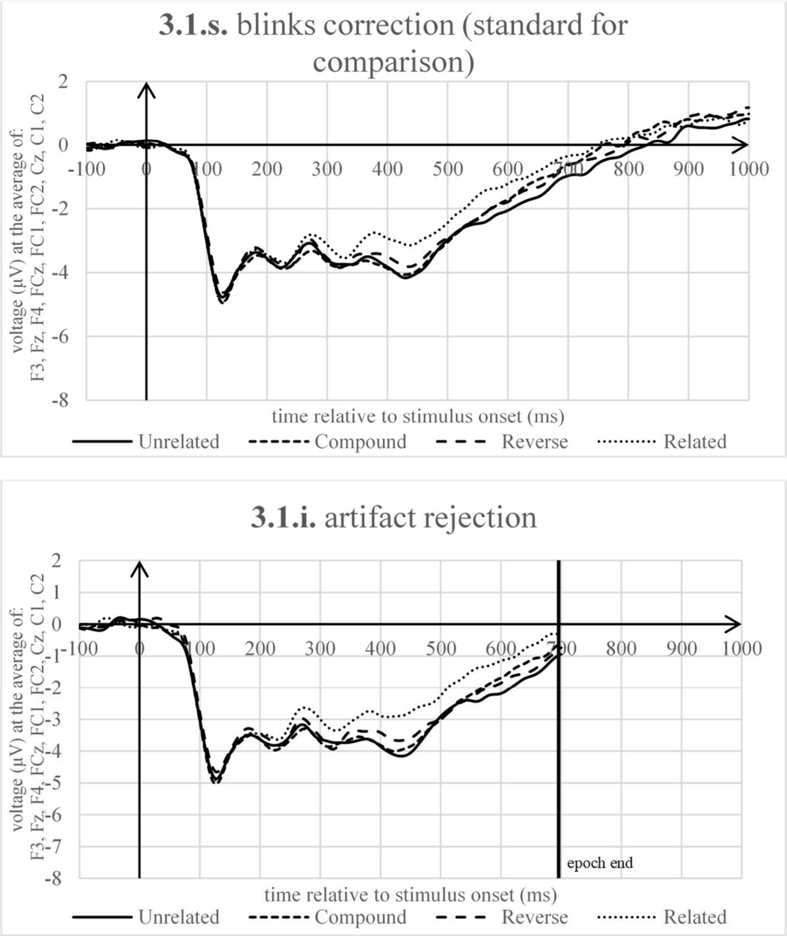


3.1.h. 350 ms



3.1. Grand average ERP waveforms

Artifact correction vs. rejection



Graph 3.1. Grand averages of experimental conditions obtained from all pre-processing and analysis pipelines. Graph 3.1.s. shows the ERPs resulting from the pipeline which was the same as the one reported by Boutonnet et al. (2014). All other pipelines deviate from this pipeline in one decision, and it was the standard for comparing other results. Note that variations in the N400 measurement window and amplitude measure were applied to data shown in this graph. To allow easier comparison, Graph 3.1.s. was printed on all pages of Graph 3.1.

Even though epochs started at -350 ms, all graphs show activity starting with baseline period which was used in each pre-processing and analysis pipeline. X-axes of graphs which show effects of artifact correction and variations in baseline duration extend beyond post-stimulus epoch or baseline period to allow easier visual contrasting.

When artifact rejection was used, the magnitude of the difference between conditions was larger, while the standard deviations were smaller for some conditions, and larger for others, and it was similar in case of using peak instead of the mean of the measurement window to quantify amplitudes.

Analysis 1: Changes in significance.

Results of one-way analyses of variance and effect sizes obtained from each waveform are shown in Table 3.3. The N400 effect remained statistically significant regardless of the pre-processing and analysis choices.

However, it was not equally prominent in all cases. The effect size, measured as partial eta squared, varied between 0.25 in case of replacing artifact correction with rejection, and 0.41 in case of prolonging baseline duration to 200 ms and re-referencing data to the mastoid reference. It was $\eta^2 = 0.36$ when the procedure described in the article by Boutonnet et al. (2014) was followed. Interestingly, this effect size was considerably larger than the effect size reported by study authors ($\eta^2 = 0.31$) (Boutonnet et al., 2014). The effect size was the lowest when artifact correction was replaced by artifact rejection and when a very wide latency window was used for mean amplitude measurement (185–635 ms), and it was also relatively low when the measurement window was set to 300–500 ms and when the digital high-pass filter with a 1 Hz cut-off was applied.

On the other hand, the effect size increased when baseline duration was extended from 100 ms to 200 ms, and when the average reference was replaced with average mastoids. Additionally, the effect size was also relatively large when the electrodes defined by the systematic review approach were used instead of the electrode set used by Boutonnet et al. (2014).

Similarly, the p value was less than 0.001 in analyses resulting from most processing and analysis pipelines. However, the omnibus effect was significant at the p<0.01 level when high-pass filter cut-off was increased to 1.0 Hz, as well as when the analysis window was changed to 300–500 or 185–635 ms. The lowest level of significance, p<0.05 was the only threshold reached when artifact correction was replaced by artifact rejection.

Table 3.3. One-way ANOVA results and effect sizes for the experimental condition effect obtained from different processing and analysis pipelines

Methodological variation	**F (3,45)**	**p**	**η_p^2**
standard	8.50	<0.001	0.36
high-pass filter			
no digital filter	8.68	<0.001	0.37
1.0 Hz	7.42	0.002 [a]	0.33
low-pass filter			
5.5 Hz	8.23	<0.001	0.35
100 Hz	8.53	<0.001	0.36
analysis window			
300–500 ms	6.72	0.001	0.31

Methodological variation	F (3,45)	p	η_p^2
185–635 ms	5.41	0.003	0.26
398–424 ms	7.66	<0.001	0.34
electrodes			
alternative set	9.56	<0.001	0.39
reference			
mastoid	10.54	<0.001	0.41
baseline duration			
200 ms	10.27	<0.001	0.41
350 ms	8.62	<0.001	0.36
artifact approach			
rejection	4.97	0.027[a]	0.25
amplitude measure			
peak	9.01	<0.001	0.38

Note: [a] Greenhouse-Geisser correction was applied because Mauchly's test was significant. All effects are significant at the α = 0.05 level. Unadjusted degrees of freedom are shown for all analyses. η_p^2 = partial eta squared; standard = pre-processing and analysis the same as the ones described by Boutonnet et al. (2014).

Analysis 2: Interaction between condition and methodological decisions.

Table 3.4. shows two-way interactions between the experimental condition and each pre-processing and analysis decision.

Table 3.4. Two-way interactions between condition and pre-processing and analysis decisions

Methodological decision	F	df	p	η_p^2
high-pass filter cut-off	7.72	6,90	**0.002** [a] *****	0.34
low-pass filter cut-off	2.43	6,90	0.087 [a]	0.14
measurement window	3.90	9,135	0.011 [a]	0.21
electrodes choice	4.83	3,45	**0.005 ***	0.24
reference	9.81	3,45	**<0.001 ***	0.40
baseline duration	3.40	6,90	0.022 [a]	0.18
artifacts approach	0.36	3,45	0.642 [a]	0.02
amplitude measure	1.18	3,45	0.326	0.07

Note: [a] Greenhouse-Geisser correction was applied because Mauchly's test was significant. * Effects are significant at the overall α = 0.05 level after applying Šidák. Unadjusted degrees of freedom are shown for all analyses. η_p^2 = partial eta-squared, df = degrees of freedom.

Out of all interactions, three methodological decisions had a statistically significant effect on the N400 effect magnitude, after Šidák correction for multiple comparisons. These decisions were: high-pass filter cut-off, electrodes choice and reference. Differences between conditions were larger when 0.01 Hz or 0.1 Hz were used, compared to 1 Hz. They were also larger when amplitudes were measured relative to the mean mastoid reference, compared to the average reference. Finally, differences between conditions were larger when frontal, frontocentral and central electrodes were used to quantify the N400 amplitude, compared to the more distributed selection of frontal, central and posterior sites.

Of course, this does not mean that differences in mean amplitudes were comparable in other conditions. As noted in the beginning of this chapter, we did not test for equivalence because it cannot be demonstrated based on only one dataset.

Analysis 3: Monte Carlo simulations.

Graph 3.2. shows the results of Monte Carlo simulations. As it can be seen from the Graph 3.2.a, which shows overall differences, statistical power depended on the pre-processing and analysis pipeline, and some decisions were more consequential than others. With only one decision changed in each pipeline, the number of trials per condition required to exceed 80% power varied between 35-70.

Selection of the measurement window had the largest effect on statistical power, which decreased the most when the analysis window extended beyond the period where the effect was the largest (300–500 and 185–635 ms windows), and less so when a narrow analysis window was used. The decision which had the second-largest negative effect on statistical power to detect the effect of condition was increasing high-pass filter cut-off to 1 Hz. Contrarywise, lowering high-pass filter cut-off had no such effect – simulated curves were almost the same in this case.

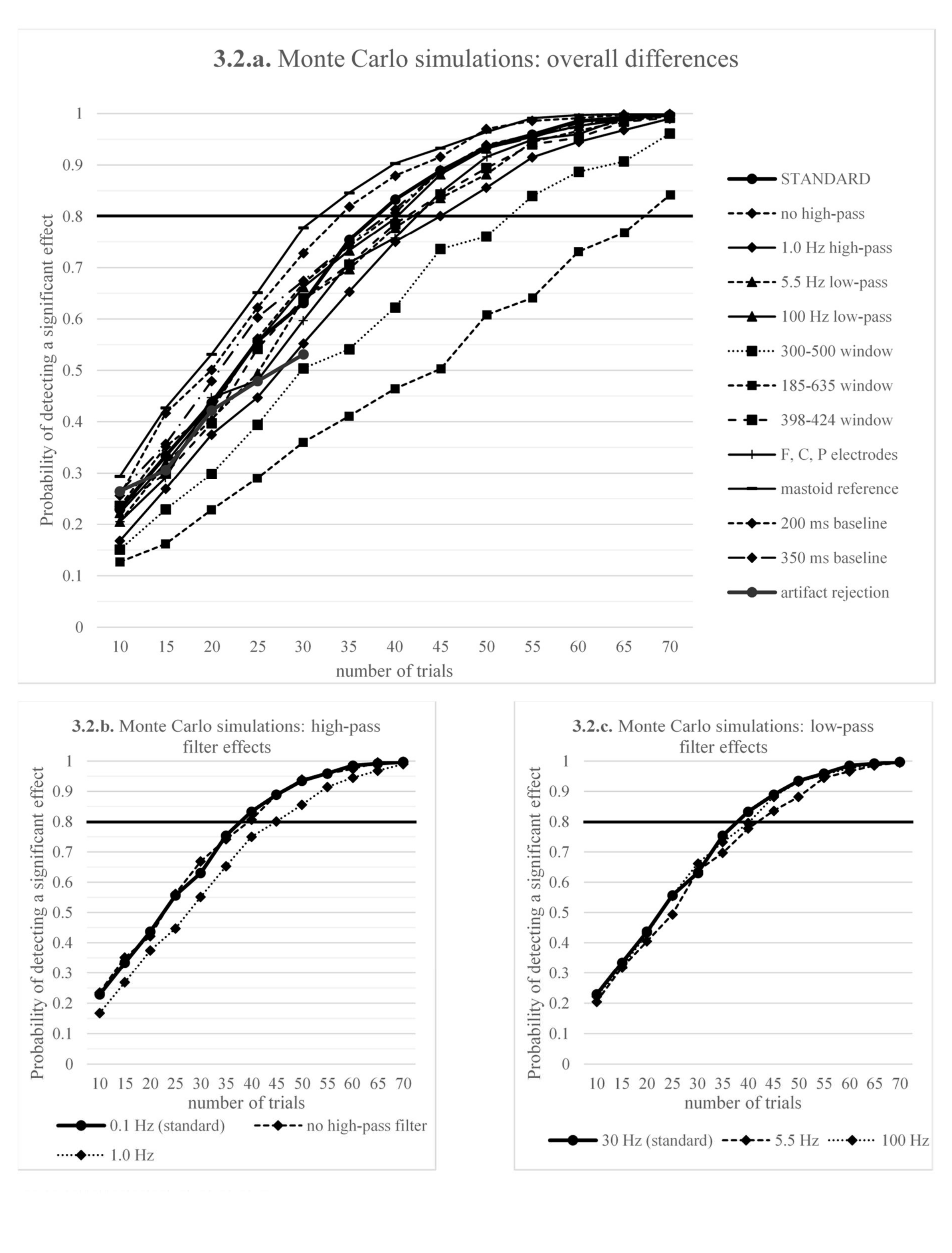
Statistical power was also somewhat reduced when a different electrode set was used, as well as when data were filtered using a low-pass filter with an extremely low cut-off (5.5 Hz). For example, with these decisions, about 5 trials per condition more were needed to reach the threshold of 80% power to detect the effect of experimental condition.

When it comes to artifact rejection, the main effect was that some participants had so few trials left that it was not possible to simulate an experiment with more than 30 trials, even though participants were shown 100 trials per condition. Interestingly, even in this trial range, statistical power was somewhat lower when artifact correction was replaced with rejection.

Finally, some variations had a positive effect on statistical power. Fewer trials were needed to achieve a given probability of obtaining a statistically significant effect when the reference was changed to average mastoids. Extending baseline duration to 200 ms had the same effect, but extending it further, to 350 ms, did not.

Discussion

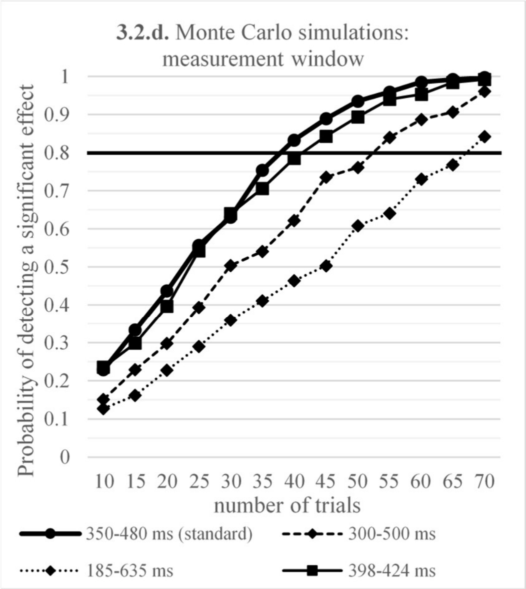
As shown in the previous section, a statistically significant N400 effect was registered consistently, regardless of the way ERP data was processed and the way the component amplitude was quantified.



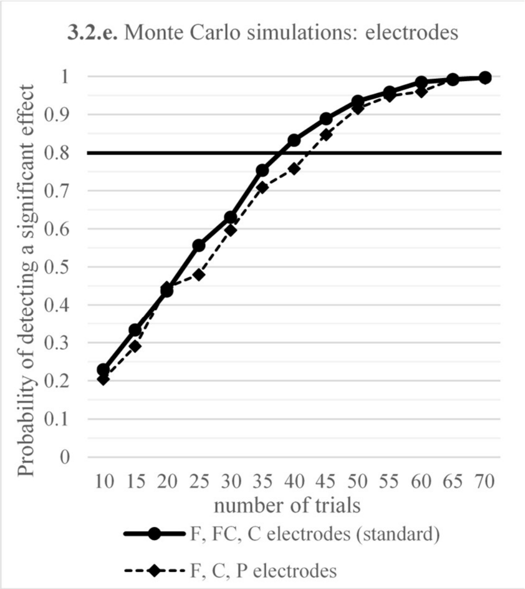
3.2.a. Monte Carlo simulations: overall differences

3.2.b. Monte Carlo simulations: high-pass filter effects

3.2.c. Monte Carlo simulations: low-pass filter effects

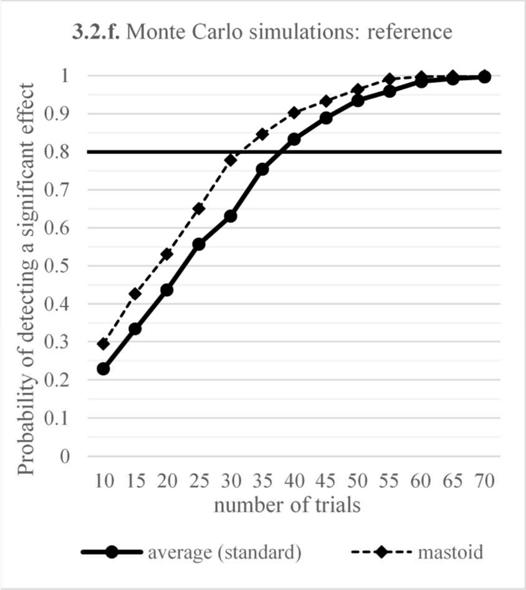


3.2.d. Monte Carlo simulations: measurement window

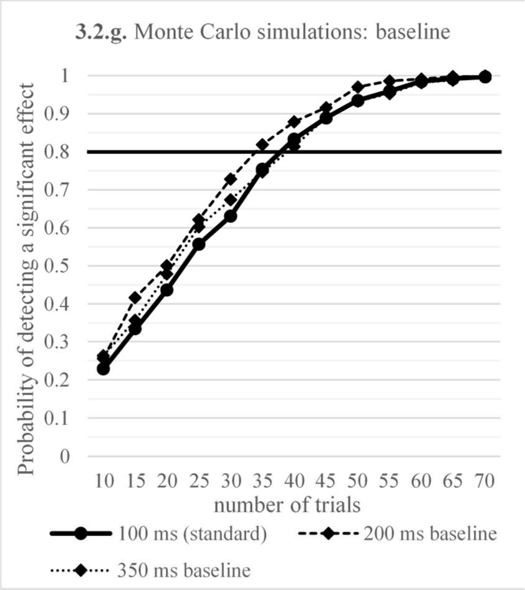


3.2.e. Monte Carlo simulations: electrodes

......



3.2.f. Monte Carlo simulations: reference



3.2.g. Monte Carlo simulations: baseline

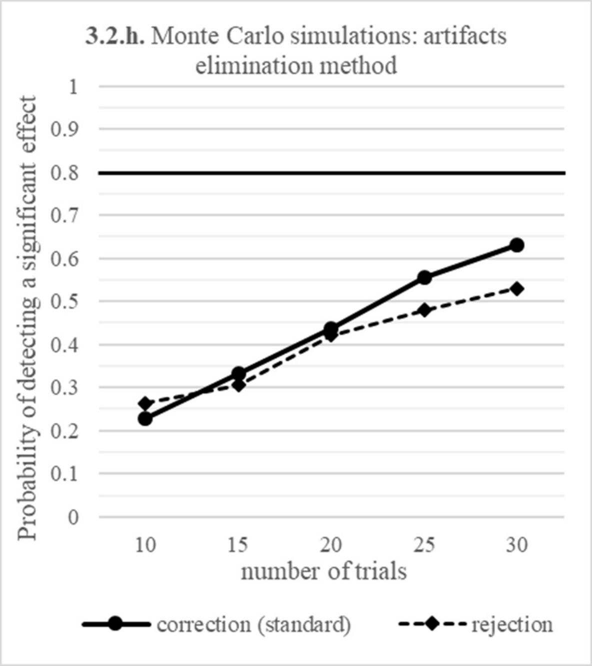


Figure 3.2. The results of Monte Carlo simulations, shown all together (3.2.a) and separately for each type of methodological variation (other graphs).

However, the effect was not completely unsusceptible to variations in methodological decisions. Most variations introduced into the pre-processing and analysis pipeline had at least some effect on the resulting waveforms and the N400 amplitudes. In fact, even when reproduced the same procedure described by Boutonnet et al. (2014) in their article, the effect size measured by partial eta squared differed by 5%.

When it comes to high-pass filters, inappropriately severe filtering distorted the shape of the waveform and reduced the magnitude of the N400 amplitude differences to a statistically significant degree. Because variability unrelated to the experimental factor was also reduced, the effect size was only somewhat lower, but statistical power to detect the effect of interest was still affected. This reduction in statistical power was in line with the effect of high-pass filter settings on the N400 amplitude demonstrated by Tanner et al. (2015). On the other hand, when digital high-pass filters were skipped altogether, leaving only the mild 0.05 Hz cut-off, we did not register any effects on the measures we observed.

Low-pass filtering did not have such a drastic effect on the N400 analyses, although the probability of obtaining a statistically significant effect with a given number of trials was mildly lowered when a low-pass filter with an extremely low cut-off was used.

Choosing a pre-stimulus baseline which was longer, but still did not include confounding pre-stimulus noise, lead to an increase in the effect size by reducing unsystematic amplitude variability, and to greater statistical power to detect this effect, even though the effect magnitude was not affected. On the other hand, due to noise in the early pre-stimulus period, waveforms slightly diverged when baseline period was extended even further. As a result, increasing baseline duration to 350 ms did not have a beneficial effect.

The magnitude of the N400 difference being similar, even a little larger, when artifact rejection was used, compared to artifact correction. Despite this, the effect size was greatly reduced, which speaks of larger within-subject variability unrelated to the experimental factor. This is unsurprising given that about a quarter of trials were eliminated per condition in this pipeline. Interestingly, statistical power seems to be reduced in this pipeline even when subsets of the same size are taken for analysis, as indicated by Monte Carlo simulations. This is unexpected given that in the pipeline in which artifact correction was used, only blinks were eliminated, and, in when artifacts rejection was performed, trials were discarded based on other large sources of noise, as well.

Like increasing baseline duration, selecting mean mastoid reference increased statistical power. However, in this case, the improvement in power was related to a statistically significant increase in the effect magnitude, produced by changes in topographic distribution which result from re-referencing (see Appendix J for topographic distributions with the two references).

The largest differences were observed when changes were made to the measurement window. The largest effect magnitude, effect size, and statistical power were observed when the 350–480 measurement window was not altered. This was not unexpected, given that this analysis window was chosen post hoc by the study's authors. However, this window was based on a circular procedure, and therefore biased towards being statistically significant (Kriegeskorte et al., 2009). The effect magnitude did not vary to a degree significant after correction for multiple comparisons when other measurement windows were taken, but the changes produced by expanding the window beyond the period of the largest effect were large enough to considerably affect partial eta squared and, consequently, statistical power. The effects of using a narrower window for mean amplitude measurement were negligible, although statistical power rose slower with the increase in the number of trials averaged together, compared to the 350–480 ms window. This decrease likely reflects not only the differences in the effect magnitude, but also in amplitude variability because fewer time points are averaged together (Luck, 2014), and it would be larger if the recordings were noisier.

On the other hand, in the analyses which we conducted here, we did not find evidence that analyses were affected by peak vs. mean amplitude choice, when all other decisions were the same as described by Boutonnet et al. (2014).

Finally, like in case of the measurement window choice, selecting a different set of electrodes did not benefit statistical power. On contrary, the waveforms on the posterior part of the scalp (P3, Pz, and P4 electrodes) which were included in the second waveform, had inverted polarity compared to the more frontal location, resulting in cancelling out some of the activity. Investigating these differences properly would require including electrode location factors in ANOVA, like in the paper by Boutonnet et al. (2014).

Summary and implications

With dozens of decisions to make after ERP data is collected and before interpreting the results, there are thousands, if not more, different ways to pre-process and analyze the same dataset. When only a subset of the main decisions was taken for analysis, and a handful of the most frequent or extreme choices were selected for each of these decisions, there was a total of 864 combinations of these decisions.

As we have seen in Chapter 2, as a result of the variety of options laid before researchers, almost each of the 132 papers included in our systematic review described at least a slightly different pre-processing and analysis strategy, even when the studies were conducted by the same team using an almost identical study design. The question we asked in this study was: does this variability matter? And the answer is, in short, yes.

None of the variations in the pre-processing and analysis strategy introduced here, even the ones clearly deviating for the standards for good practice, were enough to eliminate the omnibus N400 effect, However, as Boudewyn et al. (2018) point out, the goal of most ERP studies is not to demonstrate the presence of a component, but within- or between-subject differences in its size or timing, such as the differences examined by post hoc comparisons in the study by Boutonnet et al. (2014). Such effects are typically more subtle, and, therefore, more vulnerable to the consequences of methodological choices which were described here.

Out of decisions which were varied in this study, high-pass filter cut-off, selection of measurement latency and locations, reference, approach to eliminating artifacts, and baseline duration were all shown to affect the N400 effect magnitude, size and/or the power to detect it. Amplitude measure (peak vs. mean) and low-pass filtering did not influence the parameters which were examined.

The main conceptual limitation of this study was that it was constrained to one dataset, with its own effect magnitude and signal-to-noise ratio. Therefore, the factors which affected the results in this dataset may not have as much effect in a different study, while others, which were less relevant here, could stand out more in a different scenario. For example, in a dataset more affected by line noise, the effects of amplitude measure and low-pass filtering, which had the least impact in this dataset, would likely be more prominent. Additionally, we did not test for interactions between different methodological decisions – doing that would increase the already high complexity of the design and reduce statistical power to detect effects of interest. Finally, we examined effect magnitudes and signal-to-noise ratio, but another aspect worth examining are measures of the overall noise level in measurements, such as root mean square (RMS) of amplitudes in individual trials (see Kappenman & Luck, 2010).

On technical side, we briefly mentioned that the ERPs on all locations, including HEOG channel, had similar basic shapes, which could indicate presence of noise which could not be eliminated using the methods we pre-registered and used. This does not affect the decisions we were interested in, because we aimed to compare outcomes of a set of pre-determined variations in methodological decisions. Nevertheless, further examination could eliminate any doubts about the nature of this observation.

04 General discussion

Starting from the garden of forking paths in ERP data pre-processing and analysis, which branches out into thousands of unique pathways for each study, we set out to examine how consistent researchers are when making choices along this route, how thoroughly they document these decisions, and whether the results of their studies could depend on the turns they make at crossroads.

These issues were explored in two studies. In the first study, methodology trends in ERP research were examined using a systematic review approach, with the aim to answer the first two questions outlined above. In the second study, data from an existing, published study were used to assess how the variability in basic processing and analysis decisions which is found in the existing literature could influence experimental effects, with the goal to improve our understanding of the third question. Due to the diversity of ERP methodology, we focused on a narrow category of ERP studies, those investigating the picture-evoked N400 in healthy neurotypical adults.

The results can be summarized into the following key points.

Replicating methods: Despite the common notion that scientific papers should provide enough information to allow independent replication attempt, a researcher who would attempt to replicate picture-evoked N400 studies, and likely other ERP studies as well, would have to resort to contacting authors of the study, because relevant methodological details are frequently omitted from publications in the field. Moreover, even when information is provided, the researcher cannot be entirely sure of it. We have found evidence of conflicting information between different parts of a text and citations which do not support the claims by study authors, and the discrepancies between the paper and dataset found in Study 2 show that the information described in the paper can be incorrect due to oversight of key details or a slip in reporting.

Replicating results: Replicating results based on information available in publications can be a challenge, because subtle deviations from the method can be enough to influence the resulting ERPs. As seen in Study 2, when we reproduced the same procedure described by Boutonnet et al. (2014), on the same dataset and after asking for additional information, the effect size measured by partial eta squared differed by 5%. This is not enough to affect conclusions about the presence or absence of a large component, but it could influence more subtle differences which are often in focus of ERP studies, so ERP replication attempts should be carried out with great care and attention to detail.

Robustness of results to variations in methodological decisions: In addition to the sensitivity of ERP data to subtle variations in data pre-processing and analysis while attempting to repeat the same procedure, Study 1 has shown that the variability in procedures is far from subtle. While none of the variations selected for Study 2 managed to wipe out the N400 effect, some of the factors affected the magnitude of the effect, effect size and statistical power to detect it, demonstrating that it is not irrelevant which route one takes in the garden of ERP methods forking paths.

Deviations from guidelines for good practice: Additionally, Study 1 revealed that the guidelines for best practice are not always followed and allowed cautioning against some of the most frequent deviations, such as biased statistical analyses, inappropriate filter cut-off settings, and baseline issues.

Guidelines for a priori methodological decisions resulting from this dissertation: In addition to cautioning against the consequences of methodological choices based on the outcomes of Study 2, Study 1 allowed providing guidance for a priori selection of the measurement latency (300–500 ms) and electrode locations (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4) in picture-evoked N400 experiments. Study 2 has demonstrated that these choices may not be optimal for measurement of the N400 component, but any data-dependent decisions would require implementing appropriate corrections for the Type I error probability, thus also reducing statistical power.

***

Like the studies reviewed here, this one is not without its limitations. In the first place, its narrow focus on one modality of one component, as well as analyses applied to one dataset in Study 2, call for verifying these results and extending their generalizability by studying other subfields of ERP research and other, both simulated and real, datasets. In other words, this journey has, hopefully, only begun.

05 References

American Electroencephalographic Society. (1991). Guidelines for Standard Electrode Position Nomenclature. *Journal of Clinical Neurophysiology*, *8*(2), 200–202. https://doi.org/10.1097/00004691-199104000-00007

Baetens, K., Van der Cruyssen, L., Vandekerckhove, M., & Van Overwalle, F. (2014). ERP correlates of script chronology violations. *Brain and Cognition*, *91*, 113–122. https://doi.org/10.1016/j.bandc.2014.09.005

Balconi, M., & Pozzoli, U. (2005). Comprehending semantic and grammatical violations in Italian. N400 and P600 comparison with visual and auditory stimuli. *Journal of Psycholinguistic Research*, *34*(1), 71–98.

Balconi, M., & Vitaloni, S. (2014). N400 Effect When a Semantic Anomaly is Detected in Action Representation. A Source Localization Analysis. *Journal of Clinical Neurophysiology*, *31*(1), 58–64. https://doi.org/10.1097/WNP.0000000000000017

Barrett, S. E., & Rugg, M. D. (1989). Event-related potentials and the semantic matching of faces. *Neuropsychologia*, *27*(7), 913–922. https://doi.org/10.1016/0028-3932(89)90067-5

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition*, *14*(2), 201–212. https://doi.org/10.1016/0278-2626(90)90029-N

Barrett, S. E., Rugg, M. D., & Perrett, D. I. (1988). Event-related potentials and the matching of familiar and unfamiliar faces. *Neuropsychologia*, *26*(1), 105–117. https://doi.org/http://dx.doi.org/10.1016/0028-3932(88)90034-6

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Bensafi, M., Pierson, A., Rouby, C., Farget, V., Bertrand, B., Vigouroux, M., … Holley, A. (2002). Modulation of visual event-related potentials by emotional olfactory stimuli. *Neurophysiologie Clinique/Clinical Neurophysiology*, *32*(6), 335–342. https://doi.org/10.1016/S0987-7053(02)00337-4

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355. https://doi.org/10.1016/0013-4694(85)90008-2

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*. https://doi.org/10.1016/0013-4694(94)90094-9

Blackford, T., Holcomb, P. J., Grainger, J., & Kuperberg, G. R. (2012). A funny thing happened on the way to articulation: N400 attenuation despite behavioral interference in picture naming. *Cognition*, *123*(1), 84–99. https://doi.org/10.1016/j.cognition.2011.12.007

Bobes, M. A., Valdes-Sosa, M., & Olivares, E. I. (1994). An ERP Study of Expectancy Violation in Face Perception. *Brain and Cognition*, *26*(1), 1–22. https://doi.org/10.1006/brcg.1994.1039

Boldini, A., Algarabel, S., Ibanez, A., & Bajo, M. T. (2008). Perceptual and semantic familiarity in recognition memory: an event-related potential study. *Neuroreport*, *19*(3), 305–308. https://doi.org/10.1097/WNR.0b013e3282f4cf73

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. https://doi.org/10.1111/psyp.13049

Bouten, S., Pantecouteau, H., & Debruille, J. B. (2018). Looking for effects of qualia on event-related brain potentials of close others in search for a cause of the similarity of qualia assumed across individuals. *F1000Research*, *3*, 316. https://doi.org/10.12688/f1000research.5977.3

Boutonnet, B., McClain, R., & Thierry, G. (2014). Compound words prompt arbitrary semantic associations in conceptual memory. *Frontiers in Psychology*, *5*, 222. https://doi.org/10.3389/fpsyg.2014.00222

Bramão, I., Francisco, A., Inácio, F., Faísca, L., Reis, A., & Petersson, K. M. (2012). Electrophysiological evidence for colour effects on the naming of colour diagnostic and noncolour diagnostic objects. *Visual Cognition*, *20*(10), 1164–1185. https://doi.org/10.1080/13506285.2012.739215

Butler, D. L., Mattingley, J. B., Cunnington, R., & Suddendorf, T. (2013). Different Neural Processes Accompany Self-Recognition in Photographs Across the Lifespan: An ERP Study Using Dizygotic Twins. *PLoS ONE*, *8*(9), e72586. https://doi.org/10.1371/journal.pone.0072586

Cansino, S., Hernández-Ramos, E., & Trejo-Morales, P. (2012). Neural correlates of source memory retrieval in young, middle-aged and elderly adults. *Biological Psychology*, *90*(1), 33–49. https://doi.org/10.1016/j.biopsycho.2012.02.004

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. https://doi.org/10.1016/J.NEUROIMAGE.2012.07.004

Castle, P. ., Van Toller, S., & Milligan, G. . (2000). The effect of odour priming on cortical EEG and visual ERP responses. *International Journal of Psychophysiology*, *36*(2), 123–131. https://doi.org/10.1016/S0167-8760(99)00106-3

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, *65*(1), 1–38. https://doi.org/10.1016/j.cogpsych.2012.01.003

Cook, E. W., & Miller, G. A. (1992). Digital Filtering: Background and Tutorial for Psychophysiologists. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.1992.tb01709.x

Cooper, T. J., Harvey, M., Lavidor, M., & Schweinberger, S. R. (2007). Hemispheric asymmetries in image-specific and abstractive priming of famous faces: Evidence from reaction times and event-related brain potentials. *Neuropsychologia*, *45*(13), 2910–2921. https://doi.org/10.1016/j.neuropsychologia.2007.06.005

Dale, A. M. (1994). Source localization and spatial discriminant analysis of event-related potentials: Linear approaches (p. 175). La Jolla: University of California – San Diego.

Davis, H., Davis, P. A., Loomis, A. L., Hervey, E. N., & Hobart, G. (1939). Electrical reactions of the human brain to auditory stimulation during sleep. *Journal of Neurophysiology*, *2*, 500–514.

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of Neurophysiology*, *2*, 494–499.

Debruille, J. B., Pineda, J., & Renault, B. (1996). N400-like potentials elicited by faces and knowledge inhibition. *Brain Research. Cognitive Brain Research*, *4*(2), 133–144. https://doi.org/10.1016/0926-6410(96)00032-8

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Concerns with Nieuwland et al. (2017). San Diego: University of California. Retrieved from http://kutaslab.ucsd.edu/pdfs/FinalDUK17Comment9LabStudy.pdf

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*. https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, *34*(4), 1443–1449. https://doi.org/10.1016/j.neuroimage.2006.11.004

Demiral, Ş. B., Malcolm, G. L., & Henderson, J. M. (2012). ERP correlates of spatially incongruent object identification during scene viewing: Contextual expectancy versus simultaneous processing. *Neuropsychologia*, *50*(7), 1271–1285. https://doi.org/10.1016/j.neuropsychologia.2012.02.011

Diéguez-Risco, T., Aguado, L., Albert, J., & Hinojosa, J. A. (2013). Faces in context: Modulation of expression processing by situational information. *Social Neuroscience*, *8*(6), 601–620. https://doi.org/10.1080/17470919.2013.834842

Dien, J., & Santuzzi, A. M. (2005). Application of Repeated Measures ANOVA to High-Density ERP Datasets: A Review and Tutorial. In T. C. Handy (Ed.), *Event-Related Potentials. A Methods Handbook* (pp. 57–81). Cambridge, MA: MIT Press. https://doi.org/10.5061/dryad.30dn3

Dominguez-Martinez, E., Parise, E., Strandvall, T., & Reid, V. M. (2015). The Fixation Distance to the Stimulus Influences ERP Quality: An EEG and Eye Tracking N400

Study. *PLOS ONE*, *10*(7), e0134339. https://doi.org/10.1371/journal.pone.0134339

Donchin, E., Callaway, E., Cooper, R., Desmedt, J. E., Goff, W. R., Hillyard, S. A., & Sutton, S. (1977). Publication criteria for studies of evoked potentials (EP) in man: Methodology and publication criteria. In J. E. Desmedt (Ed.), *Progress in clinical neurophysiology: Vol. 1. Attention, voluntary contraction and event-related cerebral potentials* (pp. 1–11). Basel, Switzerland: Karger.

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

Dyck, M., & Brodeur, M. B. (2015). ERP evidence for the influence of scene context on the recognition of ambiguous and unambiguous objects. *Neuropsychologia*, *72*, 43–51. https://doi.org/10.1016/j.neuropsychologia.2015.04.023

Eddy, M. D., & Holcomb, P. J. (2009). Electrophysiological evidence for size invariance in masked picture repetition priming. *Brain and Cognition*, *71*(3), 397–409. https://doi.org/10.1016/j.bandc.2009.05.006

Eddy, M. D., & Holcomb, P. J. (2010). The temporal dynamics of masked repetition picture priming effects: Manipulations of stimulus-onset asynchrony (SOA) and prime duration. *Brain Research*, *1340*, 24–39. https://doi.org/10.1016/j.brainres.2010.04.024

Eddy, M. D., & Holcomb, P. J. (2011). Invariance to rotation in depth measured by masked repetition priming is dependent on prime duration. *Brain Research*, *1424*, 38–52. https://doi.org/10.1016/j.brainres.2011.09.036

Eddy, M. D., Schmid, A., & Holcomb, P. J. (2006). Masked repetition priming and event-related brain potentials: A new approach for tracking the time-course of object perception. *Psychophysiology*, *43*(6), 564–568. https://doi.org/10.1111/j.1469-8986.2006.00455.x

Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology*, *111*(4), 694–705. https://doi.org/10.1016/S1388-2457(99)00285-0

Elbert, T., Lutzenberger, W., Rockstroh, B., & Birbaumer, N. (1985). Removal of ocular artifacts from the EEG — A biophysical approach to the EOG. *Electroencephalography and Clinical Neurophysiology*, *60*(5), 455–463. https://doi.org/10.1016/0013-4694(85)91020-X

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: electrophysiological investigations of picture processing in the two cerebral hemispheres. *Neuropsychologia*, *40*(7), 730–747. https://doi.org/10.1016/S0028-3932(01)00193-2

Friedman, D. (1990). Cognitive Event-Related Potential Components During Continuous Recognition Memory for Pictures. *Psychophysiology*, *27*(2), 136–148. https://doi.org/10.1111/j.1469-8986.1990.tb00365.x

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123–144. https://doi.org/10.1016/S0926-6410(02)00244-6

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The Search for "Common Sense": An Electrophysiological Study of the Comprehension of Words and Pictures in Reading. *Journal of Cognitive Neuroscience*, *8*(2), 89–106. https://doi.org/10.1162/jocn.1996.8.2.89

Gao, C., Hermiller, M. S., Voss, J. L., & Guo, C. (2015). Basic perceptual changes that alter meaning and neural correlates of recognition memory. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00049

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gierych, E., Milner, R., & Michalski, A. (2005). ERP Responses to Smile-Provoking Pictures. *Journal of Psychophysiology*, *19*(2), 77–90. https://doi.org/10.1027/0269-8803.19.2.77

Giglio, A. C. A., Minati, L., & Boggio, P. S. (2013). Throwing the banana away and keeping the peel: Neuroelectric responses to unexpected but physically feasible action endings. *Brain Research*, *1532*, 56–62. https://doi.org/10.1016/j.brainres.2013.08.017

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Grave de Peralta Menendez, R., Gonzalez Andino, S., Lantz, G., Michel, C. M., & Landis, T. (2001). Noninvasive localization of electromagnetic epileptic activity. I. Method descriptions and simulations. *Brain Topography*, *14*(2), 131–137. https://doi.org/10.1023/A:1012944913650

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. https://doi.org/10.1007/BF02289823

Grigor, J. (1999). The Effect of Odour Priming on Long Latency Visual Evoked Potentials of Matching and Mismatching Objects. *Chemical Senses*, *24*(2), 137–144. https://doi.org/10.1093/chemse/24.2.137

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). 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

Gui, P., Ku, Y., Li, L., Li, X., Bodner, M., Lenz, F. A., … Zhou, Y.-D. (2017). Neural correlates of visuo-tactile crossmodal paired-associate learning and memory in humans. *Neuroscience*, *362*, 181–195. https://doi.org/10.1016/j.neuroscience.2017.08.035

Gunter, T. C., & Bach, P. (2004). Communicating hands: ERPs elicited by meaningful symbolic hand postures. *Neuroscience Letters*, *372*(1–2), 52–56.

https://doi.org/10.1016/j.neulet.2004.09.011

Hamm, J. P., Johnson, B. W., & Kirk, I. J. (2002). Comparison of the N300 and N400 ERPs to picture stimuli in congruent and incongruent contexts. *Clinical Neurophysiology*, *113*(8), 1339–1350. https://doi.org/10.1016/S1388-2457(02)00161-X

Handy, T. C. (2005). *Event-related potentials: A methods handbook*. MIT Press.

Harris, J. D., Cutmore, T. R. H., O'Gorman, J., Finnigan, S., & Shum, D. (2009). Neurophysiological indices of perceptual object priming in the absence of explicit recognition memory. *International Journal of Psychophysiology*, *71*(2), 132–141. https://doi.org/10.1016/j.ijpsycho.2008.08.005

Hauk, O. (2004). Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data. *NeuroImage*, *21*(4), 1612–1621. https://doi.org/10.1016/j.neuroimage.2003.12.018

Herring, D. R., Taylor, J. H., White, K. R., & Crites, S. L. (2011). Electrophysiological responses to evaluative priming: The LPP is sensitive to incongruity. *Emotion*, *11*(4), 794–806. https://doi.org/10.1037/a0022804

Hirschfeld, G., Feldker, K., & Zwitserlood, P. (2012). Listening to "flying ducks": Individual differences in sentence-picture verification investigated with ERPs. *Psychophysiology*, *49*(3), 312–321. https://doi.org/10.1111/j.1469-8986.2011.01315.x

Hirschfeld, G., Jansma, B., Bölte, J., & Zwitserlood, P. (2008). Interference and facilitation in overt speech production investigated with event-related potentials. *NeuroReport*, *19*(12), 1227–1230. https://doi.org/10.1097/WNR.0b013e328309ecd1

Holcomb, P. J., & McPherson, W. B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition*, *24*(2), 259–276. https://doi.org/10.1006/brcg.1994.1014

Hoogeveen, H. R., Jolij, J., Ter Horst, G. J., & Lorist, M. M. (2016). Brain Potentials Highlight Stronger Implicit Food Memory for Taste than Health and Context Associations. *PLOS ONE*, *11*(5), e0154128. https://doi.org/10.1371/journal.pone.0154128

Hoormann, J., Falkenstein, M., Schwarzenau, P., & Hohnsbein, J. (1998). Methods for the quantification and statistical testing of ERP differences across conditions. *Behavior Research Methods, Instruments, & Computers*, *30*(1), 103–109. https://doi.org/10.3758/BF03209420

Hopf, J.-M., & Mangun, G. . (2000). Shifting visual attention in space: an electrophysiological analysis using high spatial resolution mapping. *Clinical Neurophysiology*, *111*(7), 1241–1257. https://doi.org/10.1016/S1388-2457(00)00313-8

Huffmeijer, R., Tops, M., Alink, L. R. A., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2011). Love withdrawal is related to heightened processing of faces with emotional expressions and incongruent emotional feedback: evidence from ERPs. *Biological Psychology*, *86*(3), 307–313.

https://doi.org/10.1016/j.biopsycho.2011.01.003

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. *Journal of Educational Statistics*, *1*(1), 69–82. https://doi.org/10.3102/10769986001001069

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Iragui, V., Kutas, M., & Salmon, D. P. (1996). Event-related brain potentials during semantic categorization in normal aging and senile dementia of the Alzheimer's type. *Electroencephalography and Clinical Neurophysiology*, *100*(5), 392–406.

JASP team. (2019). JASP 0.11.1.0 [Computer software]. Retrieved from https://jasp-stats.org/

Jemel, B., Calabria, M., Delvenne, J. F., Crommelinck, M., & Bruyer, R. (2003). Differential involvement of episodic and face representations in ERP repetition effects. *NeuroReport*, *14*(3), 525–530. https://doi.org/10.1097/01.wnr.0000057864.05120.ba

Jennings, J. R. (1987). Editorial Policy on Analyses of Variance With Repeated Measures. *Psychophysiology*, *24*(4), 474–475. https://doi.org/10.1111/j.1469-8986.1987.tb00320.x

Jennings, J. R., & Wood, C. C. (1976). The ?-Adjustment Procedure for Repeated-Measures Analyses of Variance. *Psychophysiology*, *13*(3), 277–278. https://doi.org/10.1111/j.1469-8986.1976.tb00116.x

Jordan, T. R., & Thomas, S. M. (1999). Memory for normal and distorted pictures: Modulation of the ERP repetition effect. *Journal of Psychophysiology*, *13*(4), 224–233. https://doi.org/10.1027//0269-8803.13.4.224

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

Junghöfer, M., Elbert, T., Tucker, D. ., & Braun, C. (1999). The polar average reference effect: a bias in estimating the head surface integral in EEG recording. *Clinical Neurophysiology*, *110*(6), 1149–1155. https://doi.org/10.1016/S1388-2457(99)00044-9

Kaczer, L., Timmer, K., Bavassi, L., & Schiller, N. O. (2015). Distinct morphological processing of recently learned compound words: An ERP study. *Brain Research*, *1629*, 309–317. https://doi.org/10.1016/j.brainres.2015.10.029

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*(5), 888–904. https://doi.org/10.1111/j.1469-8986.2010.01009.x

Kappenman, E. S., & Luck, S. J. (2011). *The Oxford Handbook of Event-Related Potential Components*. (E. S. Kappenman & S. J. Luck, Eds.), *The Oxford Handbook of Event-*

Related Potential Components. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.001.0001

Kappenman, E. S., & Luck, S. J. (2016). Best Practices for Event-Related Potential Research in Clinical Populations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., … Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. https://doi.org/10.1111/psyp.12147

Khateb, A., Pegna, A. J., Landis, T., Mouthon, M. S., & Annoni, J.-M. (2010). On the Origin of the N400 Effects: An ERP Waveform and Source Localization Analysis in Three Matching Tasks. *Brain Topography*, *23*(3), 311–320. https://doi.org/10.1007/s10548-010-0149-7

Khushaba, R. N., Greenacre, L., Al-Timemy, A., & Al-Jumaily, A. (2015). Event-related Potentials of Consumer Preferences. *Procedia Computer Science*, *76*, 68–73. https://doi.org/10.1016/j.procs.2015.12.277

Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., & Townsend, C. (2013). Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Systems with Applications*, *40*(9), 3803–3812. https://doi.org/10.1016/j.eswa.2012.12.095

Kiefer, M. (2001). Perceptual and semantic sources of category-specific effects: event-related potentials during picture and word categorization. *Memory & Cognition*, *29*(1), 100–116. https://doi.org/10.3758/BF03195745

Kiefer, M., Liegel, N., Zovko, M., & Wentura, D. (2017). Mechanisms of masked evaluative priming: Task sets modulate behavioral and electrophysiological priming for picture and words differentially. *Social Cognitive and Affective Neuroscience*, *12*(4), 596–608. https://doi.org/10.1093/scan/nsw167

Kiefer, M., Sim, E.-J., Helbig, H., & Graf, M. (2011). Tracking the Time Course of Action Priming on Object Recognition: Evidence for Fast and Slow Influences of Action on Perception. *Journal of Cognitive Neuroscience*, *23*(8), 1864–1874. https://doi.org/10.1162/jocn.2010.21543

Koester, D., & Schiller, N. O. (2008). Morphological priming in overt language production: Electrophysiological evidence from Dutch. *NeuroImage*, *42*(4), 1622–1630. https://doi.org/10.1016/j.neuroimage.2008.06.043

Kovalenko, L. Y., Chaumon, M., & Busch, N. A. (2012). A Pool of Pairs of Related Objects (POPORO) for Investigating Visual Semantic Integration: Behavioral and Electrophysiological Validation. *Brain Topography*, *25*(3), 272–284. https://doi.org/10.1007/s10548-011-0216-8

Kovic, V., Plunkett, K., & Westermann, G. (2009). Shared and/or separate representations of animate/inanimate categories: An ERP study. *Psihologija*, *42*(1), 5–26.

https://doi.org/10.2298/PSI0901005K

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

Kuipers, J. R., & Thierry, G. (2011). N400 Amplitude Reduction Correlates with an Increase in Pupil Size. *Frontiers in Human Neuroscience*, *5*, 61. https://doi.org/10.3389/fnhum.2011.00061

Küper, K., Liesefeld, A. M., & Zimmer, H. D. (2015). ERP evidence for hemispheric asymmetries in abstract but not exemplar-specific repetition priming. *Psychophysiology*, *52*(12), 1610–1619. https://doi.org/10.1111/psyp.12542

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science (New York, N.Y.)*. https://doi.org/10.1126/science.7350657

Kutas, M., & Iragui, V. (1998). The N400 in a semantic categorization task across 6 decades. *Electroencephalography and Clinical Neurophysiology*, *108*(5), 456–471.

Lensink, S. E., Verdonschot, R. G., & Schiller, N. O. (2014). Morphological priming during language switching: an ERP study. *Frontiers in Human Neuroscience*, *8*, 995. https://doi.org/10.3389/fnhum.2014.00995

Li, T.-T., & Lu, Y. (2014). The subliminal affective priming effects of faces displaying various levels of arousal: An ERP study. *Neuroscience Letters*, *583*, 148–153. https://doi.org/10.1016/j.neulet.2014.09.027

Liao, S., Su, Y., Wu, X., & Qiu, J. (2011). The Poggendorff illusion effect influenced by top-down control: evidence from an event-related brain potential study. *NeuroReport*, *22*(15), 739–743. https://doi.org/10.1097/WNR.0b013e32834ab40b

Lin, M., Wang, C., Cheng, S., & Cheng, S. (2011). An event-related potential study of semantic style-match judgments of artistic furniture. *International Journal of Psychophysiology*, *82*(2), 188–195. https://doi.org/10.1016/j.ijpsycho.2011.08.007

Liu, C., Tardif, T., Mai, X., Gehring, W. J., Simms, N., & Luo, Y.-J. (2010). What's in a name? Brain activity reveals categorization processes differ across languages. *Human Brain Mapping*, *31*(11), 1786–1801. https://doi.org/10.1002/hbm.20974

Lu, A., Xu, G., Jin, H., Mo, L., Zhang, J., & Zhang, J. X. (2010). Electrophysiological evidence for effects of color knowledge in object recognition. *Neuroscience Letters*, *469*(3), 405–410. https://doi.org/10.1016/j.neulet.2009.12.039

Luck, S. J. (2005). *An Introducation to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique* (2nd ed.). Cambridge, MA: 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

Lüdtke, J., Friedrich, C. K., De Filippis, M., & Kaup, B. (2008). Event-related Potential Correlates of Negation in a Sentence–Picture Verification Paradigm. *Journal of Cognitive Neuroscience*, *20*(8), 1355–1370. https://doi.org/10.1162/jocn.2008.20093

Maffongelli, L., Bartoli, E., Sammler, D., Kölsch, S., Campus, C., Olivier, E., … D'Ausilio, A. (2015). Distinct brain signatures of content and structure violation during action observation. *Neuropsychologia*, *75*, 30–39. https://doi.org/10.1016/j.neuropsychologia.2015.05.020

Maillard, L., Barbeau, E. J., Baumann, C., Koessler, L., Bénar, C., Chauvel, P., & Liégeois-Chauvel, C. (2011). From Perception to Recognition Memory: Time Course and Lateralization of Neural Substrates of Word and Abstract Picture Processing. *Journal of Cognitive Neuroscience*, *23*(4), 782–800. https://doi.org/10.1162/jocn.2010.21434

Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R. K., & Arciuli, J. (2017). Musicians' Online Performance during Auditory and Visual Statistical Learning Tasks. *Frontiers in Human Neuroscience*, *11*, 114. https://doi.org/10.3389/fnhum.2017.00114

Manfredi, M., Adorni, R., & Proverbio, A. M. (2014). Why do we laugh at misfortunes? An electrophysiological exploration of comic situation processing. *Neuropsychologia*, *61*, 324–334. https://doi.org/10.1016/j.neuropsychologia.2014.06.029

Mao, W., & Wang, Y. (2007). Various conflicts from ventral and dorsal streams are sequentially processed in a common system. *Experimental Brain Research*, *177*(1), 113–121. https://doi.org/10.1007/s00221-006-0651-z

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

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *62*(3), 203–208. https://doi.org/10.1016/0168-5597(85)90015-2

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, *36*(1), 53–65. https://doi.org/10.1017/S0048577299971196

Mecklinger, A. (1998). On the modularity of recognition memory for object form and spatial location: a topographic ERP analysis. *Neuropsychologia*, *36*(5), 441–460. https://doi.org/10.1016/S0028-3932(97)00128-0

Miller, G. A., Lutzenberger, W., & Elbert, T. (1991). The linked-reference issue in EEG and

ERP recording. *Journal of Psychophysiology*, *5*(3), 273–276. Retrieved from https://psycnet.apa.org/record/1992-11892-001

Mnatsakanian, E. V, & Tarkka, I. M. (2003). Matching of familiar faces and abstract patterns: behavioral and high-resolution ERP study. *International Journal of Psychophysiology*, *47*(3), 217–227. https://doi.org/10.1016/S0167-8760(02)00154-X

Mnatsakanian, E. V, & Tarkka, I. M. (2004). Familiar-face recognition and comparison: source analysis of scalp-recorded event-related potentials. *Clinical Neurophysiology*, *115*(4), 880–886. https://doi.org/10.1016/j.clinph.2003.11.027

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507–517. https://doi.org/10.1016/j.neuropsychologia.2009.10.011

Mudrik, L., Shalgi, S., Lamy, D., & Deouell, L. Y. (2014). Synchronous contextual irregularities affect early scene processing: Replication and extension. *Neuropsychologia*, *56*, 447–458. https://doi.org/10.1016/j.neuropsychologia.2014.02.020

Münte, T. F., Brack, M., Grootheer, O., Wieringa, B. M., Matzke, M., & Johannes, S. (1998). Brain potentials reveal the timing of face identity and expression judgments. *Neuroscience Research*, *30*(1), 25–34. https://doi.org/10.1016/S0168-0102(97)00118-1

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP Analyses: A Step-by-Step Tutorial Review. *Brain Topography*, *20*(4), 249–264. https://doi.org/10.1007/s10548-008-0054-5

Neumann, M. F., & Schweinberger, S. R. (2008). N250r and N400 ERP correlates of immediate famous face repetition are independent of perceptual load. *Brain Research*, *1239*, 181–190. https://doi.org/10.1016/j.brainres.2008.08.039

Nielsen-Bohlman, L., & Knight, R. T. (1995). Prefrontal alterations during memory processing in aging. *Cerebral Cortex*, *5*(6), 541–549. https://doi.org/https://doi.org/10.1093/cercor/5.6.541

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience*, *4*(1), 15–22.

https://doi.org/10.1162/jocn.1992.4.1.15

Niu, Y., Xue, C., Wang, H., Zhou, L., Zhang, J., Peng, N., & Jin, T. (2016). Event-Related Potential Study on Visual Selective Attention to Icon Navigation Bar of Digital Interface. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics (EPCE 2016)* (Vol. 9736, pp. 79–89). Cham: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-40030-3_9

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*. https://doi.org/10.1016/j.jneumeth.2010.07.015

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science (New York, N.Y.)*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Olivares, E. I., & Iglesias, J. (2010). Brain potential correlates of the “internal features advantage” in face recognition. *Biological Psychology*, *83*(2), 133–142. https://doi.org/10.1016/j.biopsycho.2009.11.011

Olivares, E. I., Iglesias, J., & Bobes, M. A. (1999). Searching for face-specific long latency ERPs: a topographic study of effects associated with mismatching features. *Cognitive Brain Research*, *7*(3), 343–356. https://doi.org/10.1016/S0926-6410(98)00038-X

Olivares, E. I., Iglesias, J., & Rodríguez-Holguín, S. (2003). Long-Latency ERPs and Recognition of Facial Identity. *Journal of Cognitive Neuroscience*, *15*(1), 136–151. https://doi.org/10.1162/089892903321107873

Olivares, E. I., Saavedra, C., Trujillo-Barreto, N. J., & Iglesias, J. (2013). Long-term information and distributed neural activation are relevant for the “internal features advantage” in face processing: Electrophysiological and source reconstruction evidence. *Cortex*, *49*(10), 2735–2747. https://doi.org/10.1016/j.cortex.2013.08.001

Onton, J., & Makeig, S. (2006). Information-based modeling of event-related brain dynamics. *Progress in Brain Research*. https://doi.org/10.1016/S0079-6123(06)59007-7

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

Ortega, R., Lopez, V., & Aboitiz, F. (2008). Voluntary modulations of attention in a semantic auditory-visual matching Task: an ERP study. *Biological Research*, *41*(4), 453–460. https://doi.org/10.4067/S0716-97602008000400010

Ortiz, M. J., Grima Murcia, M. D., & Fernandez, E. (2017). Brain processing of visual metaphors: An electrophysiological study. *Brain and Cognition*, *113*, 117–124. https://doi.org/10.1016/j.bandc.2017.01.005

Ousterhout, T. (2015). N400 congruency effects from emblematic gesture probes following sentence primes. In Szakal, A (Ed.), *INES 2015 - IEEE 19TH INTERNATIONAL*

CONFERENCE ON INTELLIGENT ENGINEERING SYSTEMS (pp. 411–415). 345 E 47TH ST, NEW YORK, NY 10017 USA: IEEE. https://doi.org/10.1109/ines.2015.7329744

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, *18*(1), 49–65. https://doi.org/10.1016/0167-8760(84)90014-X

Paz-Caballero, D., Cuetos, F., & Dobarro, A. (2006). Electrophysiological evidence for a natural/artifactual dissociation. *Brain Research*, *1067*(1), 189–200. https://doi.org/10.1016/j.brainres.2005.10.046

Perez-Abalo, M. C., Rodriguez, R., Bobes, M. A., Gutierrez, J., & Valdes-Sosa, M. (1994). Brain potentials and the availability of semantic and phonological codes over time. *Neuroreport*, *5*(16), 2173–2177. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7865770

Pergola, G., Foroni, F., Mengotti, P., Argiris, G., & Rumiati, R. I. (2017). A neural signature of food semantics is associated with body-mass index. *Biological Psychology*, *129*, 282–292. https://doi.org/10.1016/j.biopsycho.2017.09.001

Picton, T. ., & Hillyard, S. . (1972). Cephalic skin potentials in electroencephalography. *Electroencephalography and Clinical Neurophysiology*, *33*(4), 419–424. https://doi.org/10.1016/0013-4694(72)90122-8

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson Jr., R., … Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*(2), 127–152. https://doi.org/10.1111/1469-8986.3720127

Picton, T. W., & Stuss, D. T. (1980). The Component Structure of the Human Event-Related Potentials. *Progress in Brain Research*, *54*, 17–49. https://doi.org/10.1016/S0079-6123(08)61604-0

Pietrowsky, R., Kuhmann, W., Krug, R., Molle, M., Fehm, H. L., & Born, J. (1996). Event-Related Brain Potentials during Identification of Tachistoscopically Presented Pictures. *Brain and Cognition*, *32*(3), 416–428. https://doi.org/10.1006/brcg.1996.0074

Pratarelli, M. E. (1994). Semantic Processing of Pictures and Spoken Words: Evidence from Event-Related Brain Potentials. *Brain and Cognition*, *24*(1), 137–157. https://doi.org/10.1006/brcg.1994.1008

Proverbio, A. M., Azzari, R., & Adorni, R. (2013). Is there a left hemispheric asymmetry for tool affordance processing? *Neuropsychologia*, *51*(13), 2690–2701. https://doi.org/10.1016/j.neuropsychologia.2013.09.023

Proverbio, A. M., Calbi, M., Manfredi, M., & Zani, A. (2014). Comprehending Body Language and Mimics: An ERP and Neuroimaging Study on Italian Actors and Viewers. *PLoS ONE*, *9*(3), e91294. https://doi.org/10.1371/journal.pone.0091294

Proverbio, A. M., Del Zotto, M., & Zani, A. (2007). The emergence of semantic categorization in early visual processing: ERP indices of animal vs. artifact recognition. *BMC Neuroscience*, *8*(1), 24. https://doi.org/10.1186/1471-2202-8-24

Proverbio, A. M., Gabaro, V., Orlandi, A., & Zani, A. (2015). Semantic brain areas are involved in gesture comprehension: An electrical neuroimaging study. *Brain and Language*, *147*, 30–40. https://doi.org/10.1016/j.bandl.2015.05.002

Proverbio, A. M., & Riva, F. (2009). RP and N400 ERP components reflect semantic violations in visual processing of human actions. *Neuroscience Letters*, *459*(3), 142–146. https://doi.org/10.1016/j.neulet.2009.05.012

Proverbio, A. M., Riva, F., & Zani, A. (2010). When neurons do not mirror the agent's intentions: Sex differences in neural coding of goal-directed actions. *Neuropsychologia*, *48*(5), 1454–1463. https://doi.org/10.1016/j.neuropsychologia.2010.01.015

Riby, L. M., & Orme, E. (2013). A familiar pattern? Semantic memory contributes to the enhancement of visuo-spatial memories. *Brain and Cognition*, *81*(2), 215–222. https://doi.org/10.1016/j.bandc.2012.10.011

Rojas, J.-C., Contero, M., Camba, J. D., Concepcion Castellanos, M., Garcia-Gonzalez, E., & Gil-Macian, S. (2016). Design Perception: Combining Semantic Priming with Eye Tracking and Event-Related Potential (ERP) Techniques to Identify Salient Product Visual Attributes. In *Proc. ASME 2015 International Mechanical Engineering Congress and Exposition, Volume 11: Systems, Design, and Complexity*. THREE PARK AVENUE, NEW YORK, NY 10016-5990 USA: AMER SOC MECHANICAL ENGINEERS.

Saavedra, C., Iglesias, J., & Olivares, E. I. (2010). Event-Related Potentials Elicited by the Explicit and Implicit Processing of Familiarity in Faces. *Clinical EEG and Neuroscience*, *41*(1), 24–31. https://doi.org/10.1177/155005941004100107

Savic, O., Savic, A. M., & Kovic, V. (2017). Comparing the temporal dynamics of thematic and taxonomic processing using event-related potentials. *PLOS ONE*, *12*(12), e0189362. https://doi.org/10.1371/journal.pone.0189362

Schendan, H. E., & Ganis, G. (2012). Electrophysiological Potentials Reveal Cortical Mechanisms for Mental Imagery, Mental Simulation, and Grounded (Embodied) Cognition. *Frontiers in Psychology*, *3*, 329. https://doi.org/10.3389/fpsyg.2012.00329

Schendan, H. E., & Ganis, G. (2015). Top-down modulation of visual processing and knowledge after 250 ms supports object constancy of category decisions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01289

Schendan, H. E., & Kutas, M. (2003). Time Course of Processes and Representations Supporting Visual Object Identification and Memory. *Journal of Cognitive Neuroscience*, *15*(1), 111–135. https://doi.org/10.1162/089892903321107864

Schleepen, T. M. J., Markus, C. R., & Jonkman, L. M. (2014). Dissociating the effects of semantic grouping and rehearsal strategies on event-related brain potentials. *International Journal of Psychophysiology*, *94*(3), 319–328. https://doi.org/10.1016/j.ijpsycho.2014.09.007

Schweinberger, S. R., Pfütze, E.-M., & Sommer, W. (1995). Repetition priming and associative priming of face recognition: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 722–736. https://doi.org/10.1037/0278-7393.21.3.722

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A Solution for Reliable and Valid Reduction of Ocular Artifacts, Applied to the P300 ERP. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.1986.tb00696.x

Shibata, H., Gyoba, J., & Suzuki, Y. (2009). Event-related potentials during the evaluation of the appropriateness of cooperative actions. *Neuroscience Letters*, *452*(2), 189–193. https://doi.org/10.1016/j.neulet.2009.01.042

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simos, P. G., & Molfese, D. L. (1997). Event-Related Potentials in a Two-Choice Task Involving Within-Form Comparisons of Pictures and Words. *International Journal of Neuroscience*, *90*(3–4), 233–253. https://doi.org/10.3109/00207459709000641

Steffensen, S. C., Ohran, A. J., Shipp, D. N., Hales, K., Stobbs, S. H., & Fleming, D. E. (2008). Gender-selective effects of the P300 and N400 components of the visual evoked potential. *Vision Research*, *48*(7), 917–925. https://doi.org/10.1016/j.visres.2008.01.005

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.

Stuss, D. ., Picton, T. ., Cerri, A. ., Leech, E. ., & Stethem, L. . (1992). Perceptual closure and object identification: Electrophysiological responses to incomplete pictures. *Brain and Cognition*, *19*(2), 253–266. https://doi.org/10.1016/0278-2626(92)90047-P

Supp, G. G., Schlögl, A., Fiebach, C. J., Gunter, T. C., Vigliocco, G., Pfurtscheller, G., & Petsche, H. (2005). Semantic memory retrieval: cortical couplings in object recognition in the N400 window. *European Journal of Neuroscience*, *21*(4), 1139–1143. https://doi.org/10.1111/j.1460-9568.2005.03906.x

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-Related ERP Components. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford Handbook of Event-Related Potential Components*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0197

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*, 997–1009. https://doi.org/10.1111/psyp.12437

Taylor, M. J., & Baldeweg, T. (2002). Application of EEG, ERP and intracranial recordings to the investigation of cognitive functions in children. *Developmental Science*, *5*(3), 318–334. https://doi.org/10.1111/1467-7687.00372

Trenner, M. U., Schweinberger, S. R., Jentzsch, I., & Sommer, W. (2004). Face repetition effects in direct and indirect tasks: an event-related brain potentials study. *Cognitive Brain Research*, *21*(3), 388–400. https://doi.org/10.1016/j.cogbrainres.2004.06.017

Trujillo-Barreto, N. J., Aubert-Vázquez, E., & Valdés-Sosa, P. A. (2004). Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, *21*(4), 1300–1319. https://doi.org/10.1016/j.neuroimage.2003.11.008

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

Wang, R. W. Y., Kuo, H.-C., & Chuang, S.-W. (2017). Humor drawings evoked temporal and spectral EEG processes. *Social Cognitive and Affective Neuroscience*, *12*(8), 1359–1376. https://doi.org/10.1093/scan/nsx054

Wang, Yinan, & Zhang, Q. (2016). Affective Priming by Simple Geometric Shapes: Evidence from Event-related Brain Potentials. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00917

Wang, Yuping, Cui, L., Wang, H., Tian, S., & Zhang, X. (2004). The sequential processing of visual feature conjunction mismatches in the human brain. *Psychophysiology*, *41*(1), 21–29. https://doi.org/10.1111/j.1469-8986.2003.00134.x

Wang, Yuping, Tian, S., Wang, H., Cui, L., Zhang, Y., & Zhang, X. (2003). Event-related potentials evoked by multi-feature conflict under different attentive conditions. *Experimental Brain Research*, *148*(4), 451–457. https://doi.org/10.1007/s00221-002-1319-y

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, *13*(3), 363–375. https://doi.org/10.1016/S0926-6410(01)00129-X

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*(3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2003). Expecting Gender: An Event Related Brain Potential Study on the Role of Grammatical Gender in Comprehending a Line Drawing Within a Written Sentence in Spanish. *Cortex*, *39*(3), 483–508. https://doi.org/10.1016/S0010-9452(08)70260-0

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*(3), 234–245. https://doi.org/10.1016/j.bandl.2006.12.003

Wu, Y. C., & Coulson, S. (2011). Are depictive gestures like pictures? Commonalities and differences in semantic processing. *Brain and Language*, *119*(3), 184–195. https://doi.org/10.1016/j.bandl.2011.07.002

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (Or Not) During Language Processing. A Commentary On Nieuwland et al. (2017) And Delong et al. (2005). *BioRxiv*, 143750. https://doi.org/10.1101/143750

Yano, T. (1995). An event-related potential study of the effects of semantic deviations: an application of a method of sequential-part presentation. *Perceptual and Motor Skills*, *81*(3 Pt 2), 1091–1098. https://doi.org/10.2466/pms.1995.81.3f.1091

Yekutieli, D., & Benjamini, Y. (2001). The control od the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Yi, A., Chen, Z., Chang, Y., Wang, H., & Wu, L. (2018). Electrophysiological evidence of language switching for bidialectals. *NeuroReport*, *29*(3), 181–190. https://doi.org/10.1097/WNR.0000000000000950

Yovel, G., & Paller, K. A. (2004). The neural basis of the butcher-on-the-bus phenomenon: when a face seems familiar but is not remembered. *NeuroImage*, *21*(2), 789–800. https://doi.org/10.1016/j.neuroimage.2003.09.034

Yum, Y. N., Holcomb, P. J., & Grainger, J. (2011). Words and pictures: An electrophysiological investigation of domain specific processing in native Chinese and English speakers. *Neuropsychologia*, *49*(7), 1910–1922. https://doi.org/10.1016/j.neuropsychologia.2011.03.018

Zani, A., Marsili, G., Senerchia, A., Orlandi, A., Citron, F. M. M., Rizzi, E., & Proverbio, A. M. (2015). ERP signs of categorical and supra-categorical processing of visual information. *Biological Psychology*, *104*, 90–107. https://doi.org/10.1016/j.biopsycho.2014.11.012

Zhang, X., Li, T., & Zhou, X. (2008). Brain responses to facial expressions by adults with different attachment-orientations. *NeuroReport*, *19*(4), 437–441. https://doi.org/10.1097/WNR.0b013e3282f55728

Zhou, H., Fan, S., Guo, J., Ma, X., Yan, J., Qin, Y., & Zhong, N. (2015). Visual Object Categorization from Whole to Fine: Evidence from ERP. In H. Guo, Y and Friston, K and Faisal, A and Hill, S and Peng (Ed.), *Brain Informatics and Health. BIH 2015. Lecture Notes in Computer Science* (Vol. 9250, pp. 325–334). Cham: Springer. https://doi.org/10.1007/978-3-319-23344-4_32

06 Appendices

APPENDIX A: Systematic review – list of online supplementary materials

Online supplements to the systematic review presented in this thesis can be found on the project's Open Science Foundation page (https://osf.io/n426j/) and they include: (1) list of exact search strings used to search Web of Science and PubMed databases (Appendix C); (2) libraries with papers found by searching these databases and papers that were selected for analysis; (3) Codebook with information on all variables used to analyze papers (Appendix D); (4) PRISMA Checklist (Appendix B); (5) a spreadsheet with all variables and information on individual papers; (6) Microsoft Excel files with analyses and graphs.

APPENDIX B: Systematic review – PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	9
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	*not included in this dissertation*
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	8-9
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	9, 11-12
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	*not applicable to this study*
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	11-12
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	11
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix C
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	11-12
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	12-14
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Appendix D
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	*not applicable to this study*
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	14
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	*not applicable to this study*

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	*not applicable to this study*
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	14, 66
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	14-18
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Appendix A
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	*not applicable to this study*
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	*not applicable to this study*
Synthesis of results	21	Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency	19-65
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	*not applicable to this study*
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	66-69
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	69-73
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	73
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	69-73, 100-101
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Acknowledgements

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

APPENDIX C: Systematic review – List of exact search strings used to search Web of Science and PubMed databases

PUBMED	
Search	**Number of hits**
N400 visual stimuli	
(N400[All Fields] AND visual[All Fields] AND stimuli[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	191
ERP N4 visual stimuli	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND visual[All Fields] AND stimuli[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	6
N400 visually evoked potentials	
(N400[All Fields] AND visually[All Fields] AND ("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields])) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	63
ERP N4 visually evoked potentials	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND visually[All Fields] AND ("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields])) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	1
N400 drawing(s)	
(N400[All Fields] AND drawing[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	9
(N400[All Fields] AND drawings[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	14
ERP N4 drawing(s)	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND drawing[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	0
also checked: drawings – no results with either	-

PUBMED	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND drawings[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	
N400 photo(graph/y)	
((((photo[All Fields] OR photograph[All Fields]) OR ("photography"[MeSH Terms] OR "photography"[All Fields])) OR photos[All Fields]) OR ("photography"[MeSH Terms] OR "photography"[All Fields] OR "photographies"[All Fields])) AND N400[All Fields] AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	15
ERP N4 photo(graph/y)	
(((photo[All Fields] OR ("photography"[MeSH Terms] OR "photography"[All Fields])) OR photos[All Fields]) OR ("photography"[MeSH Terms] OR "photography"[All Fields] OR "photographies"[All Fields])) AND (("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	0
N400 picture(s)	
(N400[All Fields] AND picture[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	138
(N400[All Fields] AND pictures[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	123
ERP N4 picture(s)	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND picture[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	2
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND pictures[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	1
N400 image(s)	
(N400[All Fields] AND ("Image (IN)"[Journal] OR "image"[All Fields])) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	37
(N400[All Fields] AND images[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	27

PUBMED	
ERP N4 image(s)	
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND ("Image (IN)"[Journal] OR "image"[All Fields])) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	2
also checked: images, but only 1 result was returned, which was found by the "image" search already	-
(("evoked potentials"[MeSH Terms] OR ("evoked"[All Fields] AND "potentials"[All Fields]) OR "evoked potentials"[All Fields] OR "erp"[All Fields]) AND N4[All Fields] AND images[All Fields]) AND ("1980/01/01"[PDAT] : "2018/06/30"[PDAT])	

Web of Science

*manually deselected papers published after June 30th, 2018 instead of setting up a filter (search conducted on 11th July 2018. Since 1996 was the earliest year available, no time limit on that end was necessary

Search	Number of hits
N400 visual stimuli	
TOPIC: (N400 visual stimuli). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	268 (+1 published after 30th June)
ERP N4 visual stimuli	
TOPIC: (ERP N4 visual stimuli). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	8
N400 visually evoked potentials	
TOPIC: (N400 visually evoked potentials). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	9
ERP N4 visually evoked potentials	
TOPIC: (ERP N4 visually evoked potentials). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	0
N400 drawing(s)	
TOPIC: (n400 drawing). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	48
TOPIC: (n400 drawing*). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	24
ERP N4 drawing(s)	
TOPIC: (ERP N4 drawing). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	0
(also checked: drawing* – no hits in either)	-
N400 photo(graph/y)	
TOPIC: (n400 photo). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	6
TOPIC: (N400 photograph). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	15
(also checked: photos, photograph* - identical results)	-
ERP N4 photo(graph/y)	

Web of Science *manually deselected papers published after June 30th, 2018 instead of setting up a filter (search conducted on 11th July 2018. Since 1996 was the earliest year available, no time limit on that end was necessary	
TOPIC: (ERP N4 photo). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	0
TOPIC: (ERP N4 photograph). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	0
(also checked: photos, photograph* - identical results)	0
N400 picture(s)	
TOPIC: (N400 picture). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	295 (+1 published after June 2018)
TOPIC: (N400 picture*). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	294 (+1 published after June 2018)
ERP N4 picture(s)	
TOPIC: (ERP N4 picture). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	3
(also checked: picture* - identical results)	-
N400 image(s)	
TOPIC: (N400 image). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	133
TOPIC: (N400 image*). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	102
ERP N4 image(s)	
TOPIC: (ERP N4 image). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	1
TOPIC: (ERP N4 image*). Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan=All years	1

Note: Multiple searches for different versions of the same word were saved when they did not give the same result.

APPENDIX D: Systematic review – Codebook with information on all variables used to analyze papers

Table D.1. Codebook for extracting information from the papers.

No	Variable	Details	Values
1	No.	unique number code for each publication	numerical variable
2	Year published	year of publication of the study	numerical variable
3	Authors	authors' names *-used to identify the study, not in the analysis*	/
4	Study design description	brief description of the study	/
5	N per group (min)	number of participants per situation *-the smallest of them if there were several groups or different analyses* *-without participants that were excluded from analyses* *-flagged with a comment if it the smallest N was in an additional analysis, and not the main one*	numerical variable
6	N total	total number of participants *-without participants that were excluded from analyses*	numerical variable
7	Presented trials – total	total number of trials seen by one participant *-the smallest of them if there was more than one experiment* *-with fillers and without practice trials*	numerical variable
8	Presented trials – per situation (min)	number of trials per situation *-the smallest of them if the number varies from situation to situation* *-excluding fillers, since they were not analyzed*	numerical variable
9	Analyzed trials per situation – type of information	did researchers report how many trials per situation were left after artifact rejection and	0 – authors did not report number of discarded trials – reported rejections – only overall

No	Variable	Details	Values
		other procedures for discarding trials	– reported average number of rejections for each situation – reported average number of rejections for each situation along with the threshold for excluding a participant or minimum number of trials for individual participants – reported rejections – only threshold for rejection or minimum number of trials – reported rejections – overall average and threshold for rejection or minimum number of trials – separately reported rejection due to artifacts and due to behavioral response error (total rejection rate inconclusive) 2 – authors did not report discarding any trials 3 – inconclusive due to unclear wording
10	Analyzed trials per situation – report	summary of information on how many trials entered ERP analysis after rejection *-if paper included information on individual experimental situations, we reported the smallest value per situation* *-exclusion of trials due to behavioral response included if it was explicitly stated that ERP epochs were rejected on this basis – otherwise, it was assumed that ERPs were not rejected based on overt response*	word summary of the analyzed report NA – not applicable (if the authors did not report discarding any trials) 0 – authors did not report number of discarded trials

No	Variable	Details	Values
11	Analyzed trials per situation – left on average*	the smallest mean number of trials per situation that was used in ERP analyses *-if paper included information on individual experimental situations, we reported the smallest value – if not, we approximated this number with average per situation for all situations* *-if rejections were reported, we calculated the number of stimuli left using this information* *-exclusion of trials due to behavioral response included if it was explicitly stated that ERP epochs were rejected on this basis – otherwise, it was assumed that ERPs were not rejected based on overt response*	numerical variable NR – information not available
12	Analyzed trials per situation – thresholds/ minimums*	minimum number of trials per situation that was used in ERP analyses, or threshold for participant exclusion *-if rejections were reported, we calculated the number of stimuli left using this information* *-exclusion of trials due to behavioral response included if it was explicitly stated that ERP epochs were rejected on this basis – otherwise, it was assumed that ERPs were not rejected based on overt response*	numerical variable NR – information not available
13	Reported hardware	reported amplifiers and cap	0 – not reported otherwise list of the equipment the way it was reported in the paper

No	Variable	Details	Values
14	Reported software	reported software for presenting trials, recording, processing and statistics	0 – not reported otherwise list of the equipment the way it was reported in the paper
15	Hardware: cap*	cap(s) that were used for EEG recording	1 – Waveguard (ANT) 2 – Electrocap 3 – QuickCap (Neuroscan) 4 – Geodesic Sensor Net 5 – Easycap standard caps 6 – Aegis Array 7 – Emotiv 8 – BrainProducts 9 – BioSemi 10 – TMSi 11 – custom Electrocap geodesic cap
16	Hardware: amplifiers*	amplifiers that were used for EEG recording	1 – Cognitrace 2 – Neuroscan 3 – Digitimer 4 – SA instrumentation 5 – BioSemi 6 – Grass 7 – NeuroScience Series III Brain Imager 8 – Brain Products 9 – Electrical Geodesics 10 – ANT 11 – TMSi 12 – M&I 13 – Emotiv 14 – Coulbourn 15 – Neuronic 16 – Nihon Kohden 17 – Psylab
17	Hardware: other*	other equipment relevant for EEG recording (e.g. electrodes, gel, specialized screens)	0 – not reported otherwise list of the equipment the way it was reported in the paper
18	Software: presenting*	software for presenting stimuli	1 – E-prime 2 – custom made 3 – Matlab 4 – ERTS 5 – Presentation

No	Variable	Details	Values
			6 – STIM 7 – EEVoke 8 – Neurolab 9 – SuperLab 10 – ERP system 11 – StimPres 12 – InstEP
19	Software: acquisition*	software for EEG data acquisition	1 – Eemagine 2 – (Neuro)scan 3 – InstEP 4 – BioSemi 5 – NetStation 6 – Brain Vision Recorder 7 – Emotiv 8 – ActiView605-Lores 9 – EEProbe 10 – TMSi REFA 11 – custom 12 – ERP System 13 – TrackWalker
20	Software: processing*	software for EEG signal processing	1 – EEGLab (+ ERPlab) 2 – InstEP 3 – FieldTrip 4 – BESA Research 5 – Scan 6 – Brain Vision Analyzer 7 – NetStation 8 – Cartool 9 – MatLab 10 – Curry 11 – TrackWalker 12 – ERP System 13 – EEProbe
21	Software: statistics*	software for statistical analysis of ERP data	1 – SPSS 2 – R 3 - BMDP4V 4 – MatLab 5 – EEGlab 6 – SAS 7 – BESA
22	Software: other*	other relevant software (e.g. modelling or visualization of data)	0 – not reported

No	Variable	Details	Values
			otherwise list of the equipment the way it was reported in the paper
23	Impedance	scalp impedance threshold	0 - not reported 1 - below 5 2 - below 10 3 - even lower than 5 other - text for other options (**note:** high-impedance solutions are noted next to numbers)
24	Reference	reference that was used in the N400 analyses *-if re-referencing took place, the new reference is reported; otherwise, online reference is reported* *-if other references were used for different components (e.g. VPP), this is not taken into account* *-linked references: if re-referencing was described, we assume that a linked reference was obtained offline by averaging; otherwise, if recording with a linked reference was mentioned, we assume that the authors referred to physical linking*	0 – not reported 1 – mean mastoids 2 – left mastoid 3 – right mastoid 4 – mastoid, unspecified or inconclusive details 5 – mean earlobes 6 – left earlobe 7 – right earlobe 8 – earlobe, unspecified or inconclusive details 9 – average (with information about whether recording montage is known) 10 – linked mastoids 11 – linked earlobes 12 – Cz 13 – tip of the nose 14 – balanced non-cephalic reference 15 – sum of mastoids
25	Online HP filter cut-off	online high-pass filter: cut-off point in Hz *-cut-offs expressed using time constant were converted into half-power cut-off points in Hz*	numerical value (NR – not reported, DC – DC filtering)
26	Online LP filter cut-off	online low-pass filter: cut-off point in Hz	numerical value (NR – not reported)
27	Offline HP filter cut-off	offline high-pass filter: cut-off point in Hz	numerical value (NR – not reported)

No	Variable	Details	Values
28	Offline LP filter cut-off	offline low-pass filter: cut-off point in Hz	numerical value (NR – not reported)
29	HP filter roll-off	online and offline high-pass filter roll-off in dB/octave *-online and offline filtering merged into one column since this was rarely reported*	0 – not reported otherwise roll-off value is given along with information whether it refers to online or offline filtering
30	LP filter roll-off	online and offline high-pass filter roll-off in dB/octave *- merged into one column since this was rarely reported*	0 – not reported otherwise roll-off value is given along with information whether it refers to online or offline filtering
31	Filtering - unit	whether cut-off points for filtering are reported using half-amplitude or half-power values	0 – not reported 1 – half-amplitude 2 – half-power 3 – online high-pass filter described using time constant
32	Notch filter	whether notch filter was used, and which frequency was suppressed (in Hz)	0 – no notch filter reported 1 – 50 Hz 2 – 60 Hz
33	N400 amplitude	what measure of amplitude was used	1 - mean on window (single waves) 2 – peak amplitude in a window (single waves) 3 – peak amplitude in a window - difference waves 4 – algorithm measure 5 – mean amplitude in a window - difference wave 6 – peak-based window 7 – point-by-point window analysis (Hoorman et al.) 8 – window area measure; 9 – no window - each time point separately throughout epoch; 0 – not reported
34	N400 window	start and end points on the waveform within which	0 – not reported; NA – not applicable (e.g. all time points were analyzed

No	Variable	Details	Values
		amplitude was measured (in milliseconds) *- for peaks and peak-centered measures, we report the window within which the researchers searched for the maximum N400 amplitude* *- if several consecutive windows were analyzed separately, both the overall window and individual windows are reported*	separately or an analysis such as PCA was applied) otherwise start and end points of the window(s) are given
35	Window reasons	reasons for choosing this specific window	1 – arbitrary, no reason given 1.1 – quoting experience, expectations, or previous work - but no reference 2 – quoting another paper 2.1 – quoting a paper that does not mention the window that was used 3 – algorithm (data driven) 4 – visual inspection (data driven) 5 – consecutive significant tests (data-driven) 6 – ANOVAs on larger consecutive windows, significance of the effect in several consecutive windows not required NA – not applicable (entire window was analyzed separately, point by point, so there was not window selection)
36	Baseline	baseline length in milliseconds - *only papers in which (1) there is a period explicitly called baseline, or (2) where baseline correction is clearly described are considered*	numerical value, NR if not reported (**note:** marked if the baseline was not pre-stimulus, but some other solution)

No	Variable	Details	Values
		- other papers are categorized as "baseline not reported" (e.g. pre-stimulus epoch is reported, but not baseline correction, or a pre-stimulus period is shown on a graph) – pre-stimulus epoch, graph epoch and baseline do not have to be the same, and in some papers they aren't the same	
37	Post-stimulus epoch	post-stimulus epoch in milliseconds	numerical value, NR if not reported
38	Epoch & components shorter than trial and overt response	whether the epoch or components overlapped with the average overt response time or the beginning of the next trial	1 – it was shorter 2.1 – epoch and a component overlap with overt reply (not with the next stimulus) 2.2 – epoch and a component overlap with the next stimulus 2.3 – only epoch overlaps with overt reply (not with the next stimulus) 2.4 – only epoch overlaps with next stimulus 2.5 – epoch overlaps with next stimulus, components with response 3 – impossible conclude from the provided information
39	Jitter	whether the interstimulus or intertrial period was jittered	0 – no 1 – yes 2 – not applicable for this type of study 3 – durations not reported
40	Break before answer	was there a break between stimulus presentation and giving overt response	0 – no 1 – yes 2 – no overt response
41	Which artifacts	which artifacts were reported in the section on artifact removal	0 – artifact handling procedure not reported,

No	Variable	Details	Values
			1 – procedure reported, but types of artifacts not reported otherwise: artifact types are listed
42	Artifacts: action taken	rejection vs. correction	0 - not reported 1 – rejected 2 – corrected 3 – combined, multistep 4 – other (what)
43	Artifacts: definition – general approach	what approach was used to define EOG artifacts	1 – visual inspection 2 – algorithm 3 – combined 4 - pre-set numerical criterion 5 – not clear 0 – artifact procedure not reported
44	Artifacts: definition - details	what approach was used to define EOG artifacts	exact strategy described in words (for combined approaches, codes are also added, the same as for the variable: “Artifacts: definition – general approach”) 0 – it was reported that artifacts were rejected or corrected, but details were not provided NA – not applicable (no artifacts procedure was reported)
45	Correction procedures*	correction procedures that were reported in the paper	verbal description 0 – it was reported that artifact correction was used, but no details were provided NA – the authors did not report using artifact correction procedures
46	Rejection procedures*	rejection procedures that were reported in the paper	verbal description 0 – it was reported that artifact rejection was used,

No	Variable	Details	Values
			but no details were provided NA – the authors did not report using artifact rejection procedures
47	Order of operations can be assumed	whether an assumption about the order of steps in data processing can be made based on the text *-it is not necessary that the order of operations is explicitly and clearly described for a positive mark*	0 – no 1 – yes
48	Electrode choice – main analysis	reason to analyze this specific combination of electrodes in the main analysis/analyses *-if all electrodes were used, but grouped into regions of interest in some way, we considered reasons for creating these regions (and not others)*	1 – quoted a paper (planned) 2 – all electrodes used in the analysis, and not grouped in any way (data driven) 3 – reason not reported (planned) 4 – visual inspection (or most prominent effect without an exact criterion, data-driven) 5 – consecutive significant tests (data-driven) 6 – algorithm (data-driven) 7 – quoted theory or own work/experience for choice, but no reference (planned) 8 – quoted a paper, but did a similar, not the same thing as the quoted researchers (planned) 9 – smallest t value/largest effect size (data-driven) NA – not applicable, only one channel
49	Recording montage – description	how many and which scalp electrodes were recorded *-without reference sites wherever possible*	description of the montage – number of electrodes, information whether exact sites are reported and list of

No	Variable	Details	Values
		-when sites are not provided in an analyzed paper, we report the electrode counts provided in analyzed papers in case exact sites were not reported[27]	electrodes for montages with smaller numbers of electrodes
50	Recording montage – layouts*	which scalp electrodes were recorded – a code for each unique analysis montage *-without reference sites wherever possible* *-when sites are not provided in an analyzed paper, we report the electrode counts provided in analyzed papers in case exact sites were not reported*	number of electrodes with a letter code for each unique montage within a given number of electrodes (e.g. – 32a, 32b… identical montages have the same electrode count and letter code) nr letter code – exact sites are not reported
51	Recording montage – count*	number of scalp channels recorded *-without reference sites wherever possible* *-when sites are not provided in an analyzed paper, we report the electrode counts provided in analyzed papers in case exact sites were not reported*	numerical variable
52	Recording montage – locations known*	are locations of electrode sites which were recorded reported in the paper	0 – no 1 – yes
53	Analysis montage - description	which electrodes were used in the main analysis/analyses - description	number of electrodes and a short description
54	Analysis montage – count*	how many electrodes were used in the main analysis/analyses	numerical variable
55	Analysis montage – layouts*	which electrodes were used in the main analysis/analyses – a	number of electrodes with a letter code for each unique

[27] Some authors that used mastoids included them in the electrode count, while others didn't (the same goes for ground electrode). Likewise, some authors that used average reference included the +1 channel that they got by re-referencing into the final count, while others didn't. Since a lot of authors didn't report which channels they had, it was often impossible to make a conclusion about whether the reference electrode was included or not.

No	Variable	Details	Values
		code for each unique analysis montage	montage within a given number of electrodes (e.g. – 32a, 32b… identical montages have the same electrode count and letter code) nr letter code – exact sites are not reported
56	Analysis montage – which electrodes*	which electrodes were used in the main analysis/analyses – only for papers in which up to 12 electrodes were analyzed *-the analysis was restricted to analysis models with 12 or fewer channels because larger montages typically involved analyzing all or most of the recorded sites, distributed over the entire scalp, while the smaller recording and analysis montages were more frequently restrictive*	-list of electrode sites for papers in which up to 12 electrode sites were analyzed (main analysis/analyses) -no value for papers with more than 12 channels in the main analysis/analyses
57	Statistics – main analysis	main analysis used to test effects that are of the central interest to the study (typically an ANOVA model)	0 – none; otherwise: exact analysis reported
58	Main analysis – unique solutions*	which electrodes were used in the main analysis/analyses of N400 amplitude[28] – a code for each unique analysis approach	analysis montage code, with an additional number code for different analyses of the same electrodes nr letter code – exact sites are not reported, but all electrodes in the cap were analyzed and number of electrodes was reported
59	Main analysis – ROI vs. individual*	whether all or only selected electrode sites were used and	1 - selected sites in ROI 3 - selected sites - analyzed individually

[28] Latency analyses, when present, are in Statistics - Additional section, while source analyses are in Statistics - Topography section.

No	Variable	Details	Values
		whether they were grouped into regions of interest	2 - all sites in ROI 4 - all sites analyzed individually
60	Main analysis – general strategy*	general approach to analyzing data (main analysis) – different groupings of factors in ANOVA, paired tests (Wilcoxon or t test) etc.	1 – ANOVA: 2x2 2 – ANOVA: 2x3 3 – ANOVA: 3x3 4 - ANOVA on PCA 5 – ANOVA: individual ROIs as factor levels 6 – ANOVA: individual sites as factor levels 7 – ANOVA: two or more separate analyses for midline and lateral electrodes 8 - ANOVA: one ROI/site 9 - all ROIs/sites in the same ANOVA – other solutions 10 - MANOVA 11 - multiple ANOVAs – other solutions 12 - not quantified – conclusions based on visual inspection 13 - approaches based on t test or Wilcoxon
61	Main statistics – correction	whether correction for multiple comparisons was applied to the main analysis	0 – no 1 – yes 2 – not necessary
62	Statistics – additional	statistical tests that were used in addition to the main analysis (e.g. post-hoc comparisons, additional ANOVAs)	0 – none; otherwise: exact analysis is reported
63	Statistics – additional: category*	categorical variable grouping different approaches to additional analyses of the N400 data	1 – post hoc and planned pairwise comparisons 1.1 – post hoc and planned - no correction reported 1.2 – Dunnet 1.3 – Tukey 1.4 – Bonferroni 1.5 – polynomial contrasts

No	Variable	Details	Values
			1.6 – Newman Keuls 1.7 – Fisher's LSD 1.8 – FDR 1.9 – Duncan multiple range test 02 – ANOVAs to test simple effects - breaking down of main ANOVA 03 – other additional ANOVAs on amplitude 04 – correlation with behavioural and other non-ERP variables 05 – ANOVA(s) on normalized data 06 – ANOVA series on smaller windows 07 – other: MANOVA, GLM-t test combination (LIMO), ERSP, t test-permutation combination (BESA), split-half reliability, linear mixed effect modelling, single-trial analysis 08 – Shapiro-Wilk test of distribution normality 09 – peak or onset latency analyses 10 – corrected post hoc ANOVAs 11 – comparison with other components - ANOVA, correlations to determine scalp similarity, etc. 12 – ANCOVA
64	Number of (M)AN(C)OVAs on N400 amplitude*	total count of ANOVAs, MANOVAs and ANCOVAs that were conducted on N400 amplitude without correction for multiple comparisons *-one (M)AN(C)OVA was considered one analysis even if*	numerical variable NA – ANOVA not used in the paper

No	Variable	Details	Values
		it had more than one factor (and thus more than one uncorrected p value) *-linear mixed models not included* *-if there are two or more experiments involving different analyses counts, the average number per experiment is provided*	
65	Correction for sphericity violation	which correction was applied for sphericity violation	0 – none GG – Greenhouse-Geisser HF – Huynh-Feldt NA – not applicable (correction not required – e.g. an ANOVA with only 2-level factors)
66	Statistics – topography	topographic maps and localization analyses	0 – none; otherwise: verbal description
67	Statistics – topography: category*	topographic maps and localization analyses: categorization	0 – no topographic analyses 1.a – voltage maps 1.b – voltage maps - normalized data 1.c – statistical maps - based on t-score, p-value 1.d – current source density maps 1.e – topographic map of PCA factors 1.f – voltage maps on percent scale 1.g – 3d voltage maps 2.a – LORETA 2.b – sLORETA 2.c – swLORETA 3 – ICA+clustering 4 – PCA 5 – spatial correlation analysis 6 – TANOVA 7 – LAURA 8 – distributed source modelling (Hauk, 2004)

No	Variable	Details	Values
			9 - Bayesian Model Averaging 10 – partial-directed coherence analysis
68	Other windows/regions	whether statistical tests were also used to test effects on other ERP components or time windows	0 – no 1 – yes (reported which components)
69	Number of other windows/regions*	number of other time windows that were analyzed using statistical tests	numerical variable
70	Earlier vs. later windows*	were other windows that were analyzed earlier or later than the N400 window *-multiple options are possible*	0 – no other windows E – earlier L – later OTHER – other (e.g. response-locked) NA – not applicable (analysis approach not based on windows)
71	Negative on plot	whether negative side was plotted upward or downward	1 – up 2 – down 3 – both
72	Supplemental materials available	whether the following supplemental materials are identifiable through the journal article: (1) materials – stimuli, scripts for running the study (2) data – raw data, individual epochs, averaged ERPs for individual participants, grand average ERPs for all conditions, amplitude measures, (3) analysis – detailed data processing pathway, codes for processing, analysis and figures	0 – no 1 – yes (reported which supplemental materials are available)
73	Additional comments	column to note things of interest that do not fit in other variables *e.g. if the paper contains reports on more than one*	/

No	Variable	Details	Values
		experiment, if several papers report different analyses of the same data, additional processing or statistical procedures	
74	Proceedings paper	column to flag proceedings from conferences	1 – proceedings from a conference 0 – journal article

Note: Variables marked by an asterisk were not extracted directly during data extraction procedure, but they were created post hoc to allow easier data analysis of variables containing verbal descriptions, which could not be categorized in advance.

APPENDIX E: Systematic review – Filter cut-off frequency – frequencies of all choices

Note: In most papers, information on whether cut-off frequency represented a half-amplitude or half-power point was not given. Therefore, all cut-offs had to be presented together here, but it should be noted that not all cases in the same category represent the same cut-off point.

Filters marked with an asterisk are potentially outside the recommended range of half-amplitude values (note the caveat that it is not known whether these cut-offs represent half-power or half-amplitude values). The half-amplitude thresholds are ≥ 0.3 Hz for high-pass filters (Luck, 2014; Tanner et al., 2015) and < 20 Hz (Luck, 2014) for low-pass filters.

Table E.1. Online high-pass filter cut-off frequencies.

Online high pass filters cut-off (Hz)	**Count**	**% of parent row total**
reported	104	78.79%
0.001	*1*	*0.96%*
0.01	*29*	*27.88%*
0.016	*10*	*9.62%*
0.02	*1*	*0.96%*
0.023	*1*	*0.96%*
0.03	*5*	*4.81%*
0.05	*23*	*22.12%*
0.08	*1*	*0.96%*
0.1	*12*	*11.54%*
0.15	*4*	*3.85%*
0.16	*3*	*2.88%*
*0.3**	*2*	*1.92%*
*0.5**	*1*	*0.96%*
*1.05**	*1*	*0.96%*
DC	*9*	*8.65%*
experiment 1: 0.1, exp 2 - 0.01	*1*	*0.96%*
not reported	28	21.21%
Grand Total	**132**	**100.00%**

Table E.2. Online low-pass filter cut-off frequencies.

Online low-pass filters cut-off (Hz)	**Count**	**% of parent row total**
reported	106	80.30%
20	*1*	*0.94%*
25	*1*	*0.94%*
30	*20*	*18.87%*
35	*1*	*0.94%*
39.2	*1*	*0.94%*

Online low-pass filters cut-off (Hz)	**Count**	**% of parent row total**
40	*19*	*17.92%*
50	*8*	*7.55%*
67	*1*	*0.94%*
69	*1*	*0.94%*
70	*10*	*9.43%*
75	*1*	*0.94%*
80	*2*	*1.89%*
100	*33*	*31.13%*
120	*1*	*0.94%*
125	*1*	*0.94%*
200	*2*	*1.89%*
250	*2*	*1.89%*
256	*1*	*0.94%*
not reported	26	19.70%
Grand Total	**132**	**100.00%**

Table E.3. Offline high-pass filter cut-off frequencies.

Offline high-pass filters cut-off (Hz)	**Count**	**% of parent row total**
reported	37	28.03%
0.01	*5*	*13.51%*
0.05	*1*	*2.70%*
0.1	*16*	*43.24%*
0.15	*2*	*5.41%*
0.16	*1*	*2.70%*
0.2	*1*	*2.70%*
*0.3**	*4*	*10.81%*
*0.5**	*3*	*8.11%*
*1**	*3*	*8.11%*
*0.3 after averaging**	*1*	*2.70%*
no offline filters, only online	92	69.70%
filtering not mentioned at all	3	2.27%
Grand Total	**132**	**100.00%**

Table E.4. Offline low-pass filter cut-off frequencies.

Offline low-pass filters cut-off (Hz)	**Count**	**% of parent row total**
reported	57	43.18%
5.5*	1	1.75%
10*	2	3.51%
15*	4	7.02%
16*	2	3.51%
20	7	12.28%
25	1	1.75%
30	27	47.37%
40	5	8.77%
45	3	5.26%
60	1	1.75%
80	1	1.75%
100	1	1.75%
30 + 5 for onset latency analysis	1	1.75%
experiment 1: 12; exp. 2 & 3 – 20*	1	1.75%
no offline filters, only online	72	54.55%
filtering not mentioned at all	3	2.27%
Grand Total	**132**	**100.00%**

APPENDIX F: Systematic review – Epoch durations

Table F.1. Frequencies of all epoch durations:

Post-stimulus epoch	**Count**	**%**
reported	110	83.33%
550	*1*	*0.91%*
600	*6*	*5.45%*
650	*2*	*1.82%*
700	*2*	*1.82%*
750	*1*	*0.91%*
800	*13*	*11.82%*
824	*1*	*0.91%*
850	*3*	*2.73%*
872	*1*	*0.91%*
874	*2*	*1.82%*
876	*2*	*1.82%*
900	*9*	*8.18%*
920	*3*	*2.73%*
924	*4*	*3.64%*
952	*1*	*0.91%*
1000	*24*	*21.82%*
1020	*1*	*0.91%*
1180	*1*	*0.91%*
1187	*1*	*0.91%*
1200	*11*	*10.00%*
1250	*1*	*0.91%*
1300	*1*	*0.91%*
1350	*1*	*0.91%*
1400	*2*	*1.82%*
1484	*1*	*0.91%*
1500	*4*	*3.64%*
1598	*1*	*0.91%*
1700	*1*	*0.91%*
1800	*2*	*1.82%*
1900	*1*	*0.91%*
2000	*4*	*3.64%*
2500	*1*	*0.91%*
600/1170	*1*	*0.91%*
inconclusive	5	3.79%
not reported	17	12.88%
Grand Total	**132**	**100.00%**

Table F.2. Relationship between baseline and epoch durations:

Epoch length grouped by baseline length (ms)	Count	%
100 ms baseline	37	45.12%
600*	1	2.70%
650*	1	2.70%
700*	1	2.70%
800*	3	8.11%
850*	2	5.41%
900*	7	18.92%
920*	2	5.41%
924*	3	8.11%
1000*	8	21.62%
1020*	1	2.70%
1180*	1	2.70%
1187*	1	2.70%
1200*	4	10.81%
1400*	1	2.70%
1500*	1	2.70%
148 ms baseline	3	3.66%
872*	1	33.33%
876*	2	66.67%
150 ms baseline	6	7.32%
600	1	16.67%
800*	1	16.67%
874*	2	33.33%
900*	1	16.67%
1598*	1	16.67%
200 ms baseline	32	39.02%
550	1	3.13%
600	1	3.13%
750	1	3.13%
800	5	15.63%
824	1	3.13%
1000	9	28.13%
1200*	6	18.75%
1250*	1	3.13%
1300*	1	3.13%
1400*	1	3.13%
1484*	1	3.13%
1800*	2	6.25%
2000*	2	6.25%
250 ms baseline	2	2.44%

Epoch length grouped by baseline length (ms)	Count	%
650	*1*	*50.00%*
1000	*1*	*50.00%*
300 ms baseline	1	1.22%
*1700**	*1*	*100.00%*
350 ms baseline	1	1.22%
1500	*1*	*100.00%*
Grand Total	**82**	**100.00%**

Note: Epoch durations which are paired with an at least 20% long baseline are marked with an asterisk.

APPENDIX G: Systematic review – Frequencies of using artifact correction and rejection procedures

Table G.1. Corrections:

Correction procedures	**Count**	**% parent row**
no details	8	16.67%
reported which algorithm was used	35	72.92%
a spatial filter algorithm (Dale, 1994) – reference could not be located	*2*	*5.71%*
Elbert et al. (1985)	*3*	*8.57%*
FASTER (Nolan et al., 2010)	*1*	*2.86%*
Gratton, Coles & Donchin (1983)	*7*	*20.00%*
ICA – independent component analysis (Jung et al., 2000)	*11*	*31.43%*
ICA-wavelet combination (Khushaba et al., 2013)	*1*	*2.86%*
interpolation using EEGLAB (Delorme & Makeig, 2004) if a numerical criterion met: ±100µV base-to-peak threshold (interpolation on a trial if less than 10% electrodes bad + interpolation of entire electrode if more than 40% trials bad)	*1*	*2.86%*
MSEC - multiple source eye correction (Berg & Scherg, 1994)	*1*	*2.86%*
Semlitsch et al. (Semlitsch et al., 1986)	*8*	*22.86%*
inconclusive	1	2.08%
only general description (regression-based)	4	8.33%
Grand Total	**48**	**100.00%**

Table G.2. Rejections:

Rejection procedures	**Count**	**% parent row**
all criteria reported	69	57.50%
±100µV base-to-peak threshold EOG + 60µV baseline to peak drift in other channels	*1*	*1.45%*
visual + ±100µV base-to-peak threshold	*1*	*1.45%*
±200 absolute, 100 peak-to-peak within 200ms threshold	*1*	*1.45%*
±70 µV threshold base-to-peak for excessive muscular activity, eye blinks, and eye movements; bad electrodes rejected only within epoch - either an average amplitude >200 µV or "difference average amplitude" >100 µV; bad electrodes rejected on all epoch	*1*	*1.45%*
±100 µV threshold (base-to-peak), differences beyond 100 µV within a 200ms interval (peak-to-peak), or activity below 0.5 µV over 100 ms	*2*	*2.90%*
±100µV base-to-peak amplitude threshold, at least100ms analog-digital clipping	*1*	*1.45%*

Rejection procedures	Count	% parent row
±100μV base-to-peak EOG and peak-to-peak threshold for other electrodes	*1*	*1.45%*
±100μV base-to-peak threshold	*2*	*2.90%*
±50 μV base-to-peak threshold	*1*	*1.45%*
±50μV peak-to-peak threshold for eye electrodes, 100μV peak-to-peak for other electrodes	*1*	*1.45%*
±70μV base-to-peak threshold	*1*	*1.45%*
±75μV base-to-peak threshold	*9*	*13.04%*
±80 μV base-to-peak threshold	*2*	*2.90%*
±80μV base-to-peak threshold	*1*	*1.45%*
±90 μV peak-to-peak threshold and ±120 μV for EOG signal	*1*	*1.45%*
125μV peak-to-peak threshold	*1*	*1.45%*
40μV peak-to-peak threshold	*1*	*1.45%*
50 μV peak-to-peak threshold	*6*	*8.70%*
base-to-peak EEG ±75μV + EOG ±100μV threshold + visual	*1*	*1.45%*
blinks: Fz base-to-peak ±60 μV threshold; other artifacts: ±80 μV base-to-peak on all electrodes	*1*	*1.45%*
1. continuous data: ±200 ms around: voltage step of of 100mV/s, max peak-to-peak within 200 ms: 400mV, min activity 0.25 mV/100 ms; *2. channel rejected if artifacts >25% recording* *3. epoch artifact: maximum allowed gradient: 25 μV/ms; maximal allowed difference within 200 ms (peak-to-peak): 100 μV; maximal positive and negative amplitude (base-to-peak): ±70; visual*	*1*	*1.45%*
FASTER toolbox (Nolan et al., 2010)	*1*	*1.45%*
if more than 10% of electrodes on a trial exceed ±100μV base-to-peak threshold; eye tracker thresholds (1° movement, fixation point miss by 1°, blink)	*1*	*1.45%*
individual threshold, peak-to-peak	*1*	*1.45%*
individual thresholds, base-to-peak	*3*	*4.35%*
maximum amplitude in the segment: ±200 μV base-to-peak threshold; maximum voltage step between successive sampling points: 50 μV; peak-to-peak between two sampling points within a 100-ms interval ±200μV	*1*	*1.45%*
outside of ±200μV (base-to-peak threshold) or difference max-min larger than 200μV (peak-to-peak threshold)	*1*	*1.45%*
threshold base-to-peak ±200 μV + voltage step per sampling point larger than 50 μV	*1*	*1.45%*
threshold: SD within 200ms windows > 50 μV	*1*	*1.45%*
ADJUST automated procedure (Mognon et al., 2011)	*1*	*1.45%*

Rejection procedures	Count	% parent row
variance 300 µV2 threshold	*1*	*1.45%*
VEOG peak-to-peak 60µV and HEOG 40µV threshold	*1*	*1.45%*
visual	*17*	*24.64%*
visual + (i) difference >140 µV between channels above and below the eyes, (ii) difference >55 µV between channels near the outer canthi or (iii) one or more channels exceeding an amplitude of 200 µV base-to-peak threshold	*1*	*1.45%*
visual bad block selection + ±75µV base-to-peak threshold	*1*	*1.45%*
rejection criteria not given	27	22.50%
all steps described, but details necessary to reproduce procedure missing	9	7.50%
criteria for some steps not described	4	3.33%
some or all steps marked as inconclusive	11	9.17%
Grand Total	**120**	**100.00%**

Rejection – common strategies	Count	%
visual inspection (% of rejection reports with all criteria given)	21	30.43%
numerical thresholds (% of rejection reports with all criteria given)	51	73.91%
individual thresholds (% of thresholds)	*4*	*7.84%*
peak-to-peak thresholds (% of thresholds)	*19*	*37.25%*
base-to-peak thresholds (% of thresholds)	*34*	*66.67%*

APPENDIX H: Systematic review – Electrode scores (Graph 2.6)

In order to provide guidance for deciding on the analysis montage based on previous literature, we examined which electrodes were reported in studies in which up to 12 electrode sites were analyzed. For this purpose, data on 65 experiments conducted on different samples was extracted from 58 publications. Within analysis montages used in these experiments, 74 different electrode labels were found. Frequency of using each electrode for analyzing data from the selected 65 experiments was registered, and, additionally, this information was weighted by the number of participants per condition in each experiment. After completing this, data was merged for locations that were labeled by different names in different papers (e.g. T5=P7), resulting in a total of 66 electrodes.

The following tables show the results, sorted by the number of times each electrode was included in statistical analysis, weighted by the number of participants per condition. The weighted column was the basis for the color scale, which was created in Microsoft Office 365 Excel (version 1909, www.microsoft.com). The colors attributed to each electrode are based on the range of scores and shown as cell backgrounds.

Electrode	Frequency of choice in experiments on separate datasets, in which analysis montages have 12 and fewer electrodes	Score: frequency of choice weighted by the number of participants per condition
Cz	36	595
Pz	27	468
F3	26	465
F4	26	465
Fz	25	457
P4	24	422
P3	24	422
C3	24	410
C4	23	397
O1	10	183
O2	10	183
Oz	10	174
FC1	7	166
FC2	7	166
FCz	8	156
F1	6	152
F2	6	152
F7	8	148
F8	8	148
T5/P7	7	138
T6/P8	7	138
FP1	7	107
FP2	7	107
F5	4	91

F6	4	91
CPz	5	84
CP4	5	83
CP2	4	75
FPz	4	73
T4/T8	4	69
C1	3	64
C2	3	64
T3/T7	4	62
CP1	3	62
P1/PP3	3	62
CP6/PC6	3	60
AF7	3	51
AF8	3	51
CP3	3	50
FC4	3	48
PO8	2	47
PO7	2	47
P2	2	42
CP5/PC5	2	40
TCP1	1	40
TCP2	1	40
FC6	2	38
CPP1h	1	30
CPP2h	1	30
FC3	2	28
PO5	1	27
PO6	1	27
AFP3h	1	26
AFP4h	1	26
FT8	1	20
P10	1	20
P9	1	20
TP8	1	20
FC5	1	18
PO3	1	18
PO4	1	18
AF3	1	14
AF4	1	14
FCC1h	1	14
FCC2h	1	14
POz	1	12

APPENDIX I: Systematic review – Reproducibility graph (Graph 2.7)

Table I.1. Frequencies from Graph 2.7 and additional information on what variables and their categories were used to calculate percentages.

Methodological information	**Applicable cases (f)**				**NA (f)**	**Note**
	Reported	*IP*	*NR*	*Total*		
Min N participants per condition	132	0	0	132	0	/
N participants - total	131	1	0	132	0	/
Analysis window	131	0	1	132	0	/
N trials - total	130	2	0	132	0	/
Min N trials per condition	122	2	0	124	8	/
Additional statistical analyses: broad categories	116	5	0	121	11	variable: Statistics – additional: category
Offline filter: low-pass cut-off value	57	0	3	60	72	NR refers to papers that do not mention filters at all, NA to papers that report using online filters
Jittering	125	1	6	132	0	/
Artifacts: rejection vs. correction	125	3	4	132	0	/
Amplitude measure	123	2	7	132	0	/
Offline filter: high-pass cut-off value	37	0	3	40	92	NR refers to papers that do not mention filters at all, NA to papers that report using online filters
Main statistical analysis: broad categories	122	10	0	132	0	variable: Main analysis – general strategy
Main statistical analysis: details	119	13	0	132	0	IP: 10 that could not be categorized + 2 which had inconclusive comments on the main analysis description + 1 which was inconclusive on the main analysis unique solution variable
Analysis montage	117	10	5	132	0	/
Epoch	110	5	17	132	0	/
Additional statistical analyses: details	98	23	0	121	11	variable: Statistics – additional; IP =

						descriptions which contained label “inconclusive”
Reference - all sites	106	23	3	132	0	IP: inconclusive + unspecified mastoid and earlobe references + the average reference if the recording montage was not given
Online filter: low-pass cut-off value	106	0	26	132	0	/
Online filter: high-pass cut-off value	104	0	28	132	0	/
Does epoch overlap with response or the next stimulus?	103	29	0	132	0	/
Baseline	102	1	29	132	0	/
Any information on hardware make & model	99	2	31	132	0	IP if any hardware was marked as inconclusive
Artifact correction algorithm	35	5	8	48	84	IP = inconclusive and procedures described only as "regression based"
Impedance	96	4	32	132	0	/
Artifacts: any details	95	11	26	132	0	Artifacts: definition – general approach; IP = category code 5
Electrodes choice rationale	84	0	48	132	0	NR: rationale not given + not known which sites were analyzed
Recording montage	78	10	44	132	0	IP = inconclusive + all electrodes but one shown
Artifact rejection method	69	24	27	120	12	variable: rejection procedures; reported = papers with all details, NR = the ones that don't have any, NA = the ones that do not have rejection, IP = all others
Analysis window choice rationale	69	18	45	132	0	IP = categories 2.1 and 1.1 as the only explanations (in at least one experiment, if multiple were conducted)

Any information on software	65	1	66	132	0	IP = any software is marked as inconclusive
Order of operations can be assumed	61	0	71	132	0	/
Offline filter: high-pass slope	17	0	20	37	95	NA = papers that have only online filters or do not include any information on filters
Offline filter: low-pass slope	19	0	38	57	75	NA = papers that have only online filters or do not include any information on filters
Filters: cut-off definition	28	10	94	132	0	IP = information given either only for online or for offline, although both are applied
Analyzed trials - all details	18	61	53	132	0	variable: Analyzed trials per situation - type of information (papers that had all details were categorized as reported, NR contains papers in which no information was given, and IP comprises all other categories)
Online filter: low-pass slope	8	1	123	132	0	/
Online filter: high-pass slope	4	1	119	124	8	NA: studies in which DC recording was used

Note: NR = not reported; NA = not applicable; IP = inconclusive or partial report; f = frequency

APPENDIX J: Methodological variations – ERP maps

ERP waveforms from the pre-processing pipeline described by Boutonnet et al. (2014) are shown, together with the pipeline in which the average reference was replaced with average mastoids. Because the goal was to simply demonstrate the overall appearance of the ERPs, which were similar in all conditions, one condition (Unrelated) is shown here.

“REF” channel is Cz, which served as the recording reference.

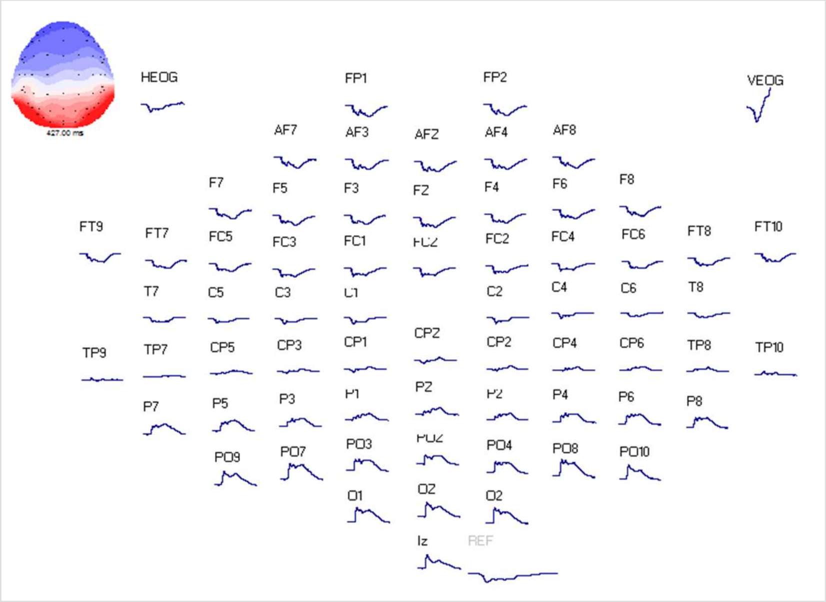


Figure J.1. ERPs – average reference (standard pre-processing pipeline)

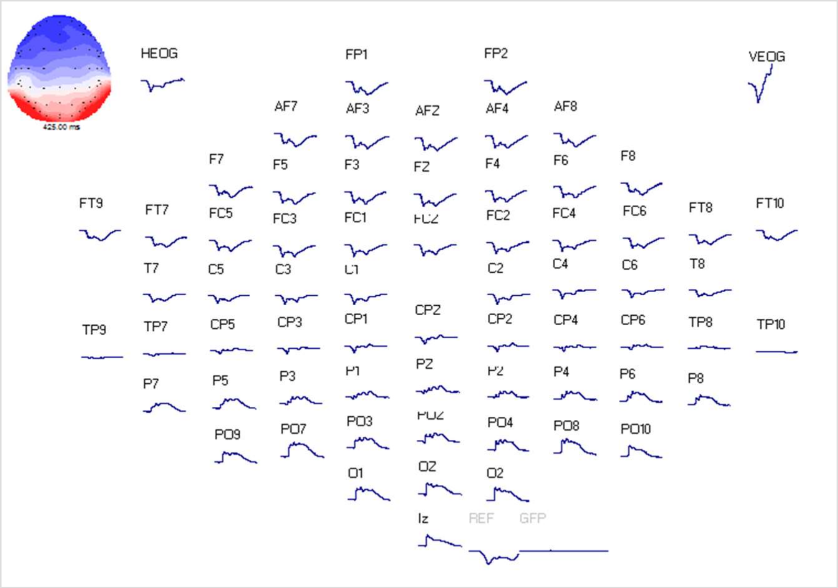


Figure J.2. ERPs – average mastoid reference

Curriculum vitae

Anđela Šoškić was born on 8th February 1992, in Belgrade. After completing the mathematics gifted student's programme in Kraljevo Grammar School in 2010, she started bachelor studies in psychology, research module, at the University of Belgrade. She completed bachelor studies in 2014, and in 2015, she obtained a master's degree in psychology, research module, at the same university. Her master thesis, "Relationship between basic personality traits and electrodermal response to aversive stimuli" was one of the top five bachelor and master theses in the 2014/15 academic year at the Foundation Katarina Marić contest. She started PhD studies at the University of Belgrade, under the mentorship of Prof Vanja Ković, in February 2016. As a student, she has taken part in numerous extracurricular activities. In 2016, she was awarded Foundation Professor Borislav Lorenc grant to support her visit to UC Davis, where she attended ERP Boot Camp in 2017.

Since 2016, she works as a Teaching Fellow at the University of Belgrade Teacher Education Faculty, where she gives seminars in developmental and educational psychology, and volunteers as a Teaching Assistant at the University of Belgrade Faculty of Philosophy. Her research interests include research methodology and developmental aspects of reading. She is a co-author of one paper („All good readers are the same, but every low-skilled reader is different: an eye-tracking study using PISA data", European Journal of Psychology of Education, 2018), and she has presented her work at several local and international conferences. Since 2018, she has been part of the research project "Fundamental cognitive processes and functions" funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

образац изјаве о ауторству

Изјава о ауторству

Име и презиме аутора Анђела Шошкић

Број индекса 4П15/08

Изјављујем

да је докторска дисертација под насловом

Evaluating ERP methodology and statistics in experiments using N400 after picture stimuli [Поређење методолошких и статистичких поступака у истраживањима потенцијала у вези са догађајем код N400 реакције на сликовну стимулацију]

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 25.11.2019.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Анђела Шошкић

Број индекса 4П15/08

Студијски програм Психологија

Наслов рада Evaluating ERP methodology and statistics in experiments using N400 after picture stimuli

Ментор Проф. др Вања Ковић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 25.11.2019.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Evaluating ERP methodology and statistics in experiments using N400 after picture stimuli [Поређење методол. и статистичких поступака у истраживањима потенцијала у вези са догађајем код N400 реакције на сликовну стимулацију]

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

[✓ 1. Ауторство (CC BY)]

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 25.11.2019.

____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.