Central Lancashire Online Knowledge (CLoK)

Title	Co-Registration of Eye Movements and Fixation-Related Potentials in Natural Reading: Practical Issues of Experimental Design and Data Analysis
Type	Article
URL	https://clok.uclan.ac.uk/id/eprint/35556/
DOI	https://doi.org/10.1525/collabra.18032
Date	2021
Citation	Degno, Federica, Loberg, Otto and Liversedge, Simon Paul (2021) Co-Registration of Eye Movements and Fixation-Related Potentials in Natural Reading: Practical Issues of Experimental Design and Data Analysis. Collabra: Psychology, 7 (1).
Creators	Degno, Federica, Loberg, Otto and Liversedge, Simon Paul

Cognitive Psychology

Co-Registration of Eye Movements and Fixation—Related Potentials in Natural Reading: Practical Issues of Experimental Design and Data Analysis

Federica Degno[1] [a], Otto Loberg[2], Simon P. Liversedge[1]

[1] School of Psychology and Computer Science, University of Central Lancashire, Preston, UK, [2] Department of Psychology, Bournemouth University, Bournemouth, UK

Collabra: Psychology

A growing number of studies are using co-registration of eye movement (EM) and fixation-related potential (FRP) measures to investigate reading. However, the number of co-registration experiments remains small when compared to the number of studies in the literature conducted with EMs and event-related potentials (ERPs) alone. One reason for this is the complexity of the experimental design and data analyses. The present paper is designed to support researchers who might have expertise in conducting reading experiments with EM or ERP techniques and are wishing to take their first steps towards co-registration research. The objective of this paper is threefold. First, to provide an overview of the issues that such researchers would face. Second, to provide a critical overview of the methodological approaches available to date to deal with these issues. Third, to offer an example pipeline and a full set of scripts for data preprocessing that may be adopted and adapted for one's own needs. The data preprocessing steps are based on EM data parsing via Data Viewer (SR Research), and the provided scripts are written in Matlab and R. Ultimately, with this paper we hope to encourage other researchers to run co-registration experiments to study reading and human cognition more generally.

Over the last 15 years, the interest in experiments that combine eye movement (EM) and event-related potential (ERP) measures, which for simplicity we will henceforth term 'co-registration' experiments, has grown. The possibility of observing continuous brain activity over time under natural reading conditions (i.e., where text is presented normally, available in both foveal and parafoveal vision, and participants make EMs as they process the text naturally in the absence of any secondary task) would be very useful to researchers wishing to explore the fine grain time-course of language processing and enhance understanding of visual word recognition (Sereno & Rayner, 2000, 2003).

Despite there being clear value in the use of co-registration as a methodological approach, it remains the case that there are only a small number of laboratories that are actively utilising this technique. Probably as a consequence of this, there is only a very limited number of published papers using this approach to study reading (see Degno et al., 2019b for a review). In our view, one of the main reasons why research in this area remains limited is because there are several significant challenges that researchers face in relation to experimental design and data analysis when using this approach.

In the present paper, we will discuss the major issues researchers encounter in co-registration reading experiments and provide an overview of what might be considered and how these issues may be dealt with. In addition, we will present a pipeline, and our implementation of the preprocessing steps of the pipeline in two scripts, one written in R and the other written in Matlab (https://osf.io/62bqx/?view-only=d28689a07d7a43bc9f8060bdf2a506a6). Whilst, to date, there exist online resources for specific sub-components of the preprocessing steps of the pipeline (e.g., see EEGLAB at https://sccn.ucsd.edu/wiki/EEGLAB#The_EEGLAB_Tutorial_Outline; Delorme & Makeig, 2004; ERPLAB at https://erpinfo.org/resources; Lopez-Calderon & Luck, 2014) and information is available in relation to specific aspects of co-registration experimentation (Dimigen, 2020; e.g., Dimigen et al., 2011; Ehinger & Dimigen, 2019; Nikolaev et al., 2016), there is no published guidance, nor implementation, of the co-registration experimental procedure as a whole.

As can be seen in Figure 1, the pipeline we present includes issues related to experimental design and data collection, separate preprocessing of EM and EEG data, matching of the two data streams, estimation of FRP measures, as well as statistical analyses of co-registered data. By discussing each step of the pipeline and providing a full set of scripts for data preprocessing, we consider that other researchers might ponder the decisions to take and adapt the associated parts of the scripts for their own needs. The order of the analysis steps within the pipeline may differ from laboratory to laboratory, and we fully acknowledge that there may be alternative solutions and scripts available, and we expect that alternative routines and software will be developed in the future. However, we are confident that the series of decisions as a whole will be faced by any researcher wishing to engage in co-registration reading experimentation. To this extent, we anticipate that the cur-
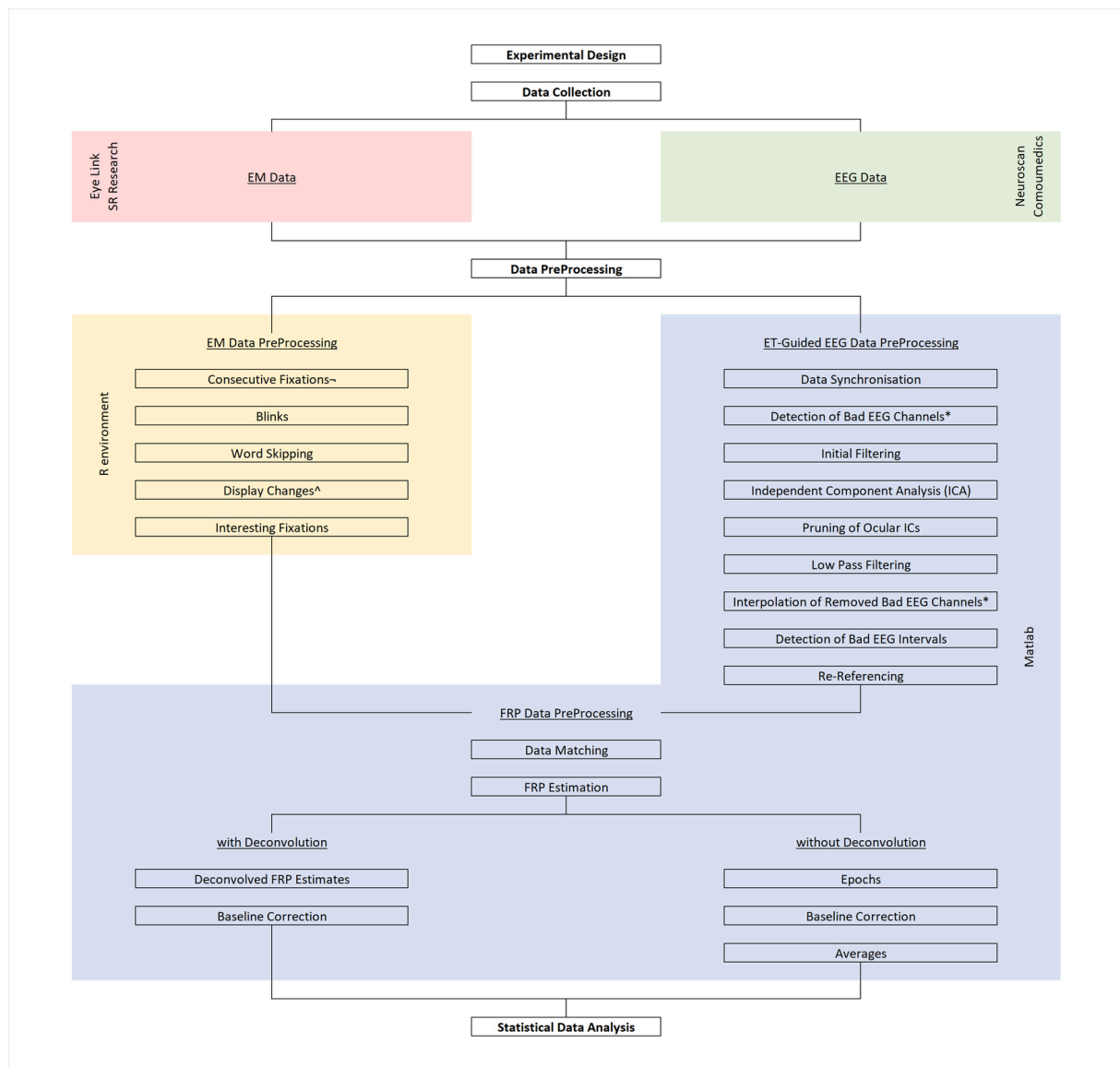
a fdegno@uclan.ac.uk

Figure 1. Example pipeline that may be adopted in order to effectively acquire and analyse data from co-registration reading experiments. In the implementation of the pipeline, EM data are assumed to be collected with an SR Research EyeLink eye tracker that allows for automatic EM parsing (e.g., detection of fixations and saccades, computation of interest area and fixation durations, etc.), whilst EEG data are assumed to be collected with Neuroscan Compumedics, or any other EEG software package that allows for the extraction of event latencies. The steps highlighted in yellow have been implemented in the R environment. The steps highlighted in blue have been implemented in Matlab. ET: eye tracking. ¬: If consecutive words are analysed (e.g., pretarget and target words). ^: If using a gaze-contingent change paradigm. *: Optional step.

rent contribution may be helpful.

The present paper is targeted at those researchers who might have an expertise in EM or EEG/ERP methods but are new to co-registration research. On one side, EM experts might find discussion of issues related to EMs relatively basic, while they might find useful the discussion of issues and procedures used in EEG/ERP research. On the other side, EEG/ERP experts might find discussion of EEG/ERP issues and procedures quite basic, as they involve standard issues that researchers face when conducting EEG/ERP research, but they might find more useful the discussion of issues that researchers face when recording and analysing EM data. Regardless of experimental background though, anyone who is interested in conducting co-registration research in reading might find the present paper useful, as parts of the pa-

per and pipeline are specific to co-registration research, and therefore new for both audiences.

We also note that the issues and pipeline we present in this paper are developed to answer research questions that relate to both foveal and parafoveal processing during reading, using gaze-contingent display change paradigms (see Rayner, 1975). Gaze-contingent display change paradigms (e.g., invisible boundary paradigm, moving window paradigm; see Rayner, 1998 for a review) allow the experimenter to make rapid changes to the visual stimulus (e.g., a sentence) a participant is looking at based on their eye position. In turn, this allows researchers to understand the extent and type of information that readers process in parafoveal and peripheral vision during reading. Depending on the researchers' theoretical questions, all or only some

of the issues we cover will need to be considered. For example, if only foveal processing is to be investigated, some decisions related to parafoveal processing and gaze-contingent display change paradigms might not be necessary (e.g., early and late display changes in the EM data preprocessing). We will point out whether each issue should be considered in relation to foveal and/or parafoveal processing and/or different paradigms in each instance.

1. Experimental Design

1.1 Experimental Approaches

The main advantage of conducting co-registration experiments to investigate reading is that it allows researchers to examine the neural correlates of written language comprehension as it occurs naturally. Approximations of natural reading (as distinct from natural reading itself) are conditions in which (1) participants are presented with sequences of words, such that the visual and linguistic content is processed in foveal and, to some extent, parafoveal vision on any particular fixation, and (2) participants make saccadic EMs to sample the words in the way they do when reading text outside the laboratory. In the published literature to date, two experimental approaches have been adopted that fall within this definition: saccadic word-list reading and sentence reading. We will consider pros and cons of each approach below.

1.1.1. Saccadic Word-List Reading

In a saccadic word-list reading task, a list of unrelated words (usually about 5 or so) is presented horizontally on the screen and participants are required to move their eyes from left to right, in order to silently read each word and then either judge their semantic category (e.g., Dimigen et al., 2012; Kornrumpf et al., 2016; Niefind & Dimigen, 2016) or perform a recognition task (e.g., Hutzler et al., 2007, 2013). It may be argued that there are advantages to this paradigm: it allows for each word in the list to be manipulated experimentally; it removes the necessity for the inclusion of short function words in the list (thereby reducing potentially problematic word skipping behavior); it (arguably) requires the participant to fixate each word at least once in order to perform the task; it avoids confounds such as predictability effects from sentential context. However, despite these positive characteristics, in our view, this approach is significantly limited in its delivery of insight into natural written language comprehension. Readers cannot form any meaningful interpretation beyond the lexical or perhaps syntactic representation of each individual word. Critically therefore, this approach does not allow the participant to read naturally in the absence of a secondary task (e.g., a memory task or a semantic category decision task) and might lead to unnatural linguistic processing (e.g., different influences of word frequency were observed in saccadic word-list reading, Niefind & Dimigen, 2016 and in sentence reading, Degno et al., 2019a; Kretzschmar et al., 2015).

1.1.2. Sentence Reading

In sentence reading tasks, typically, a one-line sentence (Degno et al., 2019a, 2019b; Dimigen et al., 2011; Kretzschmar et al., 2009, 2015; Loberg et al., 2018, 2019; Metzner et al., 2015, 2017; Mirault et al., 2020; Weiss et al., 2016), or longer passages of text (Henderson et al., 2013), is presented on the screen, and participants are required to read

each sentence to then answer comprehension questions. By presenting grammatical text (minimally a single sentence, and potentially longer texts), readers can form a meaningful interpretation of the text beyond isolated word meaning, and all aspects of EM behaviour during reading (e.g., skipping), as well as the participant's task (i.e., reading for comprehension without a secondary task) are natural. Certainly, when adopting this approach, the experimental design and analyses become a little more complex. First, the stimuli need to be carefully designed to avoid an influence of extraneous linguistic variables (e.g., predictability and plausibility, as well as the length of pre-target, target and post-target words). Unless these variables are themselves to be manipulated, they are generally controlled to prevent the effect of interest becoming contaminated by their influence. Second, it is likely that by using sentences or passages of text a larger loss of data may occur (e.g., readers are likely to skip some of the words, especially function words), compared to word-list reading. Nevertheless, and for this very reason, the sentence reading approach reflects the cognitive processes that underpin natural reading, and therefore, we consider it the desirable approach to use if the goal is to investigate *natural* reading.

1.2. Stimuli Construction

The appropriate number of trials per conditions depends on both the number of participants and the size of the effect being investigated (e.g., Boudewyn et al., 2017). However, it is fundamental to consider a number of points beyond this when constructing the stimuli and developing the experiment.

First, it is important to note that in co-registration experiments data loss is higher than the data loss observed when only EMs or EEG/ERPs are recorded. Indeed, as can be seen in Figure 1, co-registered data are cleaned according to both EM and EEG criteria, and only those data that pass both cleaning procedures are matched and used for final FRP analyses. Second, it is important to take into account that although increasing the number of trials per condition usually improves the statistical power of the study, to optimise the signal-to-noise ratio, a balance between quality of the recorded data (i.e., recording data from subjects that are not fatigued) and time spent testing participants should also be achieved (see Boudewyn et al., 2017 for a discussion). It is straightforward to create large numbers of stimuli in a saccadic word-list reading task, as each word in the list can be manipulated and analysed, each being independent from the others in the list. This significantly increases the potential number of observations per trial. However, constructing a large number of stimuli for a natural sentence reading experiment is more difficult. Typically, in sentence reading studies a single target word is experimentally manipulated in each sentence. In these circumstances, a large number of sentences needs to be created because the number of target words per stimulus is reduced, and therefore, it is more likely that readers might skip the target in a trial. Also, it is more likely that observations on the target word might be lost during the cleaning procedures. In our experience, to minimise such difficulties, whenever possible, more than one target word should be included in each experimental sentence, thereby increasing the number of observations obtainable from each trial (though of course, the greater the number of target words in a sentence, the harder it is to construct text that is natural, and to some extent, meaningful).

With respect to the design of the stimuli themselves, en-

suring that the targets words, and for sentence reading experiments also pre-target and post-target words, are at least 4 characters long will reduce the skipping rate on those interest areas (Rayner, 1979), and thus, reduce data loss. In addition, as mentioned above, extraneous variables such as cloze probability and plausibility of sentence should be controlled (unless these effects themselves are to be investigated).

1.3. Experimental Procedures

A number of good practices can be adopted to minimise the disruption to the EEG signal caused by simultaneous recording of EMs, and at the same time to take advantage of the simultaneous use of both methods. First, procedures can be adopted to reduce the number of oculomotor artifacts that lead to increased data loss. In particular, it is important to reduce the number of blinks that participants make while reading, as otherwise there is great possibility that a blink might occur on an interest area, and thus, the associated observations would be lost. To achieve this, it can be helpful to present participants with a 'blink screen' after each trial (i.e., a screen presentation with the word "BLINK" presented centrally to which participants are instructed they must blink). This procedure substantially reduces the frequency with which participants blink during experimental trials, which in turn reduces the probability of participants blinking on the critical interest areas.

Second, as is standard in EM research, it is good to use a chin rest and head restraint to minimise head movements. However, to reduce any pressure that might impact EEG recordings at frontal sites, it may be helpful to modify the head restraint making it padded to cushion the forehead, or to remove the forehead rest and rely only on the chinrest.

Finally, in our experiments, we have found it to be beneficial to ask participants to fixate a cross on the right hand side of the screen when they have finished reading the text, rather than to press a button on a response box, to terminate the current trial and trigger the following one. The fixation on the box is recorded via the eye tracking (ET) system, and the following trial is thereby triggered automatically. By using this procedure, it is possible to prevent the EEG signal becoming contaminated with any activity related to motoric preparations associated with the manual response required in the use of a response box (e.g., Van der Lubbe et al., 2000).

2. Data Collection

2.1. Connection

A physical electronic connection that delivers a temporally fast and accurate signal between the ET and EEG recording devices is essential to obtain meaningful co-registration data. The most common stimuli presentation software (e.g., Experiment Builder, E-Prime, Presentation, etc.), ET systems (e.g., SR Research EyeLink, Tobii Pro eye trackers, etc.) and EEG systems (e.g., Neuroscan Compumedics, Brain Vision, EGI, etc.) support such connections between different devices, although they might differ in the way in which connection is achieved (Dimigen et al., 2011). The stimulus display computer might send a TTL (Transistor-Transistor Logic) signal to both EEG and ET devices via a splitter or Y-shape cable connected to a parallel port. In such a situation, this signal will appear both in the EEG signal and EM data as an event marker. Alternatively, a message might be sent to the ET system from the stimulus display computer via an ethernet link, and a TTL signal sent to

the EEG system via a parallel port from the stimulus display computer or from the ET computer. Using this type of connection, the TTL signal will be displayed as an event marker in the EEG signal, and the message will be recorded in the EM data as a string of text. Another option is to add an analog card to the ET computer to convert the digital data into analog voltages and make a connection between the ET computer and the EEG amplifiers (with the ET analog outputs inserted as channels in the EEG headbox). The analog voltages will then need to be converted back into digital data by the EEG recording software. Using this type of connection, the eye positions (e.g., horizontal and vertical position, pupil size) will be directly lined up with the EEG signal and displayed as separate ET channels. Although this method allows for online synchronisation, it brings a series of significant disadvantages (e.g., added noise during the digital/analog signal conversion, employment of EEG channels as ET channels; see Dimigen et al., 2011 for a discussion), which undermine its usability in co-registration experiments. Other types of connections, specific to the systems in use in one's laboratory, might also be needed (e.g., a combination of ethernet links and fibre optic links for EGI systems). Thus, the specific type of connection that is required will depend on the stimuli presentation software, ET and EEG systems, and physical connections that may be available.

2.2. Synchronization

Another important aspect of co-registration is ensuring that ET and EEG systems are temporally well synchronised with each other. To achieve this, it is critical to make sure that a sufficient number of event markers are sent simultaneously to both the ET and EEG systems, with the appropriate code and at the correct timepoint in each trial.

An event marker indicates a point in time during the recording when a specific event occurs. We suggest that the minimum number of markers in the experiment would be one marker at the beginning and one at the end of each trial. By having two markers in each trial, it is possible to compare the timing between trials and the duration of each trial in both the ET and EEG recordings (of course, these should be very similar if the experiment is set up adequately). In addition, two markers are required by the EYE-EEG extension (Dimigen et al., 2011) of EEGLAB in Matlab. This tool allows for a check that the EEG events are aligned with the ET events in the two data streams, and it calculates any synchronisation error (for details see the online tutorial at http://www2.hu-berlin.de/eyetracking-eeg/tutorial.html, in particular the sections "Basics: Connecting eye tracker & EEG", "Basics: Synchronization signals", and "Step 5: Check synchronization accuracy (via cross-correlation)"). In our experience, using this tool to examine the timing accuracy prior and during the testing period, as well as during the offline data preprocessing can be very helpful.

Different event markers can be sent during data acquisition (with the exact values depending on the number of bits available, e.g., values between 0-255 with 8-bits). It is possible to send event markers with codes that are unique to each trial and/or condition (e.g., 3 for the onset and 103 for the offset of a trial in the experimental condition 3), and/or to send the same repeated markers in correspondence with particular events (e.g., 31 for all trial onsets and 32 for all trial offsets). The EYE-EEG extension requires the same repeated code for trial onsets (e.g., 31) and offsets (e.g., 32) to conduct synchronisation checks. However, if only unique event markers are sent during the experiment it is possible

to recode them to meet the EYE-EEG requirements. Similarly, if only repeated event markers are sent during data acquisition, it is possible to retrieve trial and condition information offline. Ideally, researchers might want to send both types of event markers for easier synchronisation and trial identification. In this case, researchers might send an event marker at trial onset and one at trial offset, as well as an event marker (e.g., trial and/or condition identifier) within the period of time between these two triggers.

Finally here, we note that although it is possible to send a marker in correspondence with interest areas (e.g., onset of a fixation on a target word) during data acquisition, in this type of experiment, it is not ideal to do this. It takes a variable amount of time (e.g., on average 35 ms with the EyeLink tracker; see SR Research Experiment Builder User Manual v2.2.245, page 184) for an ET system to detect a fixation in real-time. A similar, variable delay will then be observed when sending the marker to the recordings, making it difficult to remove with precision the delay, and carrying the risk that the data are not perfectly aligned to the veridical fixation onset. Thus, to calculate the onset of an event associated with an interest area (e.g., onset of a fixation on a target word), a better alternative is to calculate this timing offline. For example, by calculating the latency between trial onset and fixation onset, it is possible to determine the timing of the first fixation onset on a target word within the trial. We consider this offline procedure to be more accurate, as the marker is immediately sent in conjunction with the onset of the trial, without any delay. This will ensure that the timing of coincident events is precise.

2.3. Recording Parameters

We consider online sampling rate and filtering to be important aspects of co-registration research that need careful consideration during data acquisition.

Sampling rate is the number of samples that are recorded per second. The online rate of sampling should be high enough to capture the known cortical activity of interest, that is related to cognitive processing and measured with EEG (roughly up to 110 Hz; e.g., Muthukumaraswamy, 2013 for overview of high frequency phenomena), and account for the Nyquist frequency (i.e., the highest known frequency content that can be reconstructed without aliasing), which corresponds to half the sampling rate (e.g., 500 Hz when recording with 1000 Hz sampling rate).

When choosing the online sampling rate, it is important to remember that if both EM and EEG data are recorded with the same sampling rate, this will ensure a perfect sample correspondence and will make matching of the two data streams easier. If the sampling rate of the two instruments is not the same, precision of the co-registered data will be determined by the least accurate device, with the number of (real, not estimated) samples for the combined and matching co-registered data to correspond to the lowest sampling rate of the two devices. Thus, researchers might choose to record both EM and EEG data with a high sampling rate (e.g., 1000 Hz), which can be downsampled offline, if computation of data processing becomes unfeasible and reduction of the resolution of the data is needed. Alternatively, researchers might favour smaller, more manageable data sets, recording data online with a lower sampling rate (e.g., 500 Hz or 256 Hz). In all cases, the limits set by the frequency of the brain activity of interest and the Nyquist frequency need to be taken into account. In addition, it is important to remember that although upsampling is mathematically possible, the added data points will be an estimation of the

activity that should have occurred during those time points, not a true observation.

Regarding online filtering, when choosing the parameters, it is important to remember that once data are filtered, they cannot be unfiltered. Thus, it is better to use more stringent filters during the offline EEG/ERP data preprocessing rather than during the online data acquisition. This said, however, it is necessary to use online anti-aliasing filters to suppress frequencies at or above the Nyquist frequency, to avoid introducing artifactual frequencies in the EEG signal (see chapter 7 of Luck, 2014 for detailed recommendations). Therefore, similarly to the sampling rate, the online filters need to be chosen considering the known cortical activity of interest and the Nyquist frequency.

3. Preprocessing of Data

The analysis steps of the pipeline we present here are not intended to be prescriptive with respect to either the parameters to adopt for each of the issues that are considered, or the order in which to make these decisions. Indeed, many of the decisions to be taken in relation to EEG/ERP/FRP preprocessing are open to debate. Thus, we refer the reader to the literature for a more exhaustive discussion of each of these issues. Here we limit ourselves to providing an overview of the issues to consider and discussing available approaches to deal with these issues when engaging in co-registration reading research. We also note that given the numerous decisions researchers need to make for the data preprocessing, specifying in advance one's own analysis plan, or at least some of the decisions a priori, for example with preregistration, could be beneficial.

The pipeline covers issues that relate to co-registration experimentation when both foveal and parafoveal processing during reading are to be investigated (see Rayner, 1998, 2009 for reviews). Thus, both pre-target and target words are considered regions of interest in the present pipeline. However, researchers might select only pre-target or only target words as regions of interest, if for example only parafoveal-on-foveal effects are examined (e.g., pre-target word only, Mirault et al., 2020), or only foveal processing is to be studied (e.g., target words only, Hutzler et al., 2007). In addition, which observations need to be retained or discarded will also vary depending on the research question at hand. In some circumstances, fixations after regressive EMs may be discarded (e.g., Degno et al., 2019a), whilst under other circumstances, these could be the focus of the experimental investigation (e.g., Metzner et al., 2017). Thus, in general, depending on the theoretical questions that are being investigated, and in turn, the analysis one is conducting, selection of interest areas and observations will vary.

3.1. PreProcessing of EM Data

3.1.1. Eye Movement Parsing

Eye-movement detection (e.g., parsing of saccades, fixations, and blinks), interest area assignment (e.g., pre-target and target words), as well as calculation of EM measures (e.g., first fixation durations) is the basis for preprocessing and analysis of EM and FRP data. It can be conducted with widely available automatized commercial software (e.g., SR Research EyeLink). Alternatively, it is possible to parse the EM data with specific algorithms (e.g., Ralf Engbert & Kliegl, 2003, R. Engbert & Mergenthaler, 2006 as implemented in EYE-EEG extension of EEGLAB, http://www2.hu-berlin.de/eyetracking-eeg, Dimigen et al., 2011), and then

assign interest areas and compute EM measures with custom scripts.

Both approaches have been used in the co-registration reading literature (SR Research automatic parsing, Degno et al., 2019a, 2019b; Henderson et al., 2013; Hutzler et al., 2013; López-Peréz et al., 2016; Mirault et al., 2020; detection with Engbert et al.'s algorithm, Dimigen et al., 2011; Dimigen et al., 2012; Kornrumpf et al., 2016; Loberg et al., 2018, Loberg et al., 2019; Metzner et al., 2015; Metzner et al., 2017; Niefind & Dimigen, 2016; detection with Stampe's (1993) algorithm, Baccino & Manunta, 2005; detection with Smeets & Hooge's (2003) algorithm, Simola et al., 2009; detection with Nyström & Holmqvist's (2010) algorithm, Weiss et al., 2016). Researchers might choose one approach over the other depending on the software they have available and on the flexibility that they need for EM parsing (e.g., if detection of very small saccades or computation of global velocity is needed). However, automatic and custom scripts should produce highly similar EM parsing results.

3.1.2. Consecutive Fixations

A typical pattern of EMs in reading includes (1) fixations, which are the periods of time (on average, approximately 250 ms) when the eyes remain relatively still to extract useful information from the text, (2) forward saccades, which are the movements the eyes make to bring new information into foveal vision (with an average saccade length of about 7-9 letter spaces for alphabetic languages like English), and (3) regressions, which are backward saccades typically occurring because of text or comprehension difficulties after approximately 10-15% of fixations (Rayner, 2009).

Thus, during natural reading, readers might make only one fixation on both pre-target and target words, or multiple fixations on the pre-target word to then fixate once or multiple times the target word. In all these cases, the reader makes consecutive fixations in first-pass reading on the pre-target and then the target word, as there is at least one fixation on the pre-target word, followed by at least one fixation on the target word. These trials truly reflect foveal and parafoveal processing associated with the regions of interest and are therefore important in relation to the theoretical investigation. In contrast, there are some instances of fixation patterns in which the pre-target and target words are fixated non-consecutively. For example, readers could make a fixation on the pre-target word, followed by a fixation on a previous interest area, or they might make a fixation on the pre-target word but then skip the target word to fixate a later interest area, or else, they might fixate the target word before making a fixation on the pre-target word. In all these cases, there are not consecutive fixations on pre-target and target words during first pass-reading.

There is evidence showing that whether readers make a progressive saccade forward or a regressive saccade backwards, this might be indicative of a correction of oculomotor error and differences in the nature of cognitive processes under each circumstance (Schotter & Rayner, 2015). In turn, these differences might produce differences in the neural correlates associated with these effects (e.g., Metzner et al., 2017). Thus, when running a co-registration reading experiment to investigate both foveal and parafoveal processing, it would be ideal to include in the analyses only those observations where there are consecutive fixations on the words of interest during first pass-reading. That is, only those observations in which readers first make a fixation on the pre-target word, immediately followed by at least one fixation on the target word, and possibly followed by at least

one fixation on the post-target word (although consecutive fixations from target to post-target word might be less of an issue if parafoveal processing or very early foveal processing of the target word is investigated) (Degno et al., 2019a, 2019b; see also Henderson et al., 2013 for inclusion of fixations preceded by a rightward saccade only, and Metzner et al., 2015 for inclusions of words that received a progressive fixation only).

3.1.3. Blinks

Blinks or vertical EMs (that cause the eyelid to move over the eyeball to some extent) reflect the upward, or downward, and the nasalward movement of the upper eyelid over the eyeball (Collewijn et al., 1985). Under normal viewing conditions, eye blinks occur every few seconds (e.g., Nakano et al., 2013) and are accompanied by suppression of visual input despite the subjective feeling of continuity of vision (Volkmann et al., 1980). Blinks represent a source of noise in the co-registration data. First, during blinks the eyelid occludes the pupil of the eyes, and therefore the visual input is suppressed during such events and readers do not obtain any new information. Second, if a blink is detected around the fixation on the target word, the preceding and/or following fixations might be corrupted (e.g., see SR Research EyeLink 1000 User Manual v1.5.0, page 110). A single fixation on the target word might be interrupted by a blink, and therefore, being split into two separate fixations (i.e., the fixation preceding the blink, and the fixation following the blink). Under these circumstances, we cannot be sure that the first fixation on the target word reflects the same cognitive processing as when no blinks occur. Indeed, if a second fixation occurs following a blink, it might be likely that further processing is necessary to process the target word. Third, the EEG signal associated with blinks is much larger in magnitude than electrical potentials generated by the brain (e.g., see Table 1 of Keren et al., 2010), producing oculomotor distortion into the EEG signal.

As mentioned earlier, presenting participants with a 'blink screen' after each trial (e.g., Degno et al., 2019a, Degno et al., 2019b; Hutzler et al., 2007; Hutzler et al., 2013) considerably reduces the amount of blinking during reading. If a blink does occur during a trial, it is possible to identify it during fixations on the critical interest areas and remove the associated observations (Baccino & Manunta, 2005; Degno et al., 2019a, Degno et al., 2019b; Dimigen et al., 2011; Dimigen et al., 2012; Henderson et al., 2013; Hutzler et al., 2007; Hutzler et al., 2013; Kornrumpf et al., 2016; Kretzschmar et al., 2015; Loberg et al., 2019; Mirault et al., 2020; Niefind & Dimigen, 2016; Simola et al., 2009; Weiss et al., 2016). The observations associated with blinks can be identified and removed from the EM dataset, in order to ensure that they will not be used inadvertently as time-locking events for the FRP data, and/or from the EEG dataset (e.g., detecting EEG intervals where the ET data are out of range, such as when a blink occurs, during the procedure for independent component analysis of the ET-guided EEG data preprocessing).

Whichever option is chosen, and regardless of the paradigm adopted and parafoveal or foveal processing being investigated, it is important to remove the noise caused by eye blinks to obtain meaningful data that represent the ongoing cognitive processes underlying reading rather than oculomotor actions.

3.1.4. Word Skipping

On average skilled readers skip 15% of content words and 65% of function words (Rayner, 2009). There is consensus that when a word is skipped it is processed, to some extent, on the fixations prior and after the skip (e.g., Drieghe et al., 2005; Pollatsek et al., 1986; Rayner et al., 2003, 2004; Reichle et al., 1998). Thus, skipping any of the interest areas during first-pass reading implies that (i) the fixation prior to the skip might include processing of the skipped word to a different extent than when the upcoming word is fixated, (ii) the fixation made after the skip might include some spillover processing, again to a greater extent than would occur when there is no skip, (iii) if the words of interest are fixated for the first time during second-pass fixations (i.e., a word is skipped and then later fixated after a regression from a word downstream in the sentence), the processing that occurs during such fixations is likely to reflect quite different cognitive processes to those that occur during first pass fixations on the word. For these reasons, skipping is considered another source of noise for co-registration data, which needs consideration.

The easiest method to reduce skipping of critical words embedded in a list or sentence is to use words (ideally pre-target, target and post-target words) that are at least four characters long (Rayner, 1979). Alternatively, or in addition, researchers might consider removing those observations where a skip occurred on the preceding, current, or following area of interest. As for consideration of consecutive fixations, skipping the post-target word might be less problematic if parafoveal processing of the target word or early foveal processing of the target word are the aspects of processing that are under investigation.

It is important to consider the skipping issue for investigations of both foveal and parafoveal processing in co-registration reading research, regardless of the paradigm being used (Degno et al., 2019a, 2019b; Dimigen et al., 2012; Metzner et al., 2015; Mirault et al., 2020). However, obviously, when skipping behavior is itself the issue of investigation, then such observations must be retained and examined in detail (e.g., Kretzschmar et al., 2015).

3.1.5. Early and Late Gaze-Contingent Display Changes

Gaze-contingent display change paradigms (see Rayner, 1998 for a review) allow the experimenter to make rapid changes to the visual display a participant is looking at contingent on their eye position. This allows us to understand the extent and type of information that readers process in parafoveal and peripheral vision during reading (or other tasks).

For these paradigms to work, it is necessary that the display change occurs very rapidly during a saccade, when visual input is suppressed (so-called *saccadic suppression*; Matin, 1974), and not during the preceding or subsequent fixation (e.g., Angele et al., 2016; Richlan et al., 2013; Slattery et al., 2011).

Early display changes might occur for several reasons, including a blink or the readers' gaze being very close to the boundary, microsaccades, sampling errors, and hooks. Hooks are usually small post-saccadic EMs that occur when the eye does not land at exactly the intended location at the end of a saccade but slightly overshoots that location and returns to it in the opposite direction (*dynamic overshoot*; Bahill et al., 1975a, 1975b; Bahill & Stark, 1975; Kapoula et al., 1986). Taking hooks as an example, if this rapid "wob-

ble" like eyeball movement (Nyström & Holmqvist, 2010) is sufficient for the gaze to temporarily transgress the boundary, this will trigger the change, and the display will be updated earlier than expected while the fixation remains slightly to the left of the boundary on the pre-boundary word. If the display change is triggered in this way, it will occur too early and the parafoveal manipulation will not be effective.

Likewise, if the display change occurs too late, foveal processing of the target word will be affected. Late display changes occur when the parafoveal manipulation (e.g., a mask) is still present when the eyes fixate on that word (e.g., when fixating the post-boundary word in a boundary paradigm experiment). Late changes are mainly caused by temporal errors in the presentation screen refresh rate. When the display update is too slow, the participant will experience the early portion of the fixation in the target region directly fixating the preview, that is, processing a stimulus they were not intended to fixate directly and this will then change to the target during the fixation. Such a situation means that the participant will be more likely to be aware of the stimulus change and that the stimulus presentation experience of the participant will not be that which was intended. Thus, early and late changes are considered a potential source of noise in the data and need to be taken into account.

To avoid early and late display changes, the ET system needs to rapidly and accurately calculate the location of the point of fixation over time and then swiftly use this information to update the display before the end of the saccade. Thus, sampling frequency should be high enough to ensure that a rapid display change is achieved. The higher the sampling rate, the faster the system will receive information regarding the location of the point of fixation, and therefore, the more rapidly this information can be used to update the display. In addition, it is desirable to use stimuli displays with a rapid refresh rate to ensure fast implementation of the display update (e.g., Richlan et al., 2013). Furthermore, it is important that the experiment is the only task being run on the computer during testing (Krantz, 2000) to ensure that the computer is as temporally responsive as possible, thereby minimising the possibility of delayed update. Finally, it is useful to check the actual latency of the display update using photosensors (e.g., Richlan et al., 2013), and offline, after data acquisition, the delay in the data (which might vary somewhat between trials), and exclude observations where updates occurred early or were slow (e.g., Wang & Inhoff, 2013).

It is important to consider this issue when any gaze-contingent display change paradigm is used and parafoveal processing is investigated in co-registration reading research (Degno et al., 2019a, 2019b; Dimigen et al., 2012; Kornrumpf et al., 2016; Niefind & Dimigen, 2016). However, when no gaze-contingent display change paradigm is adopted, this issue will not pose a problem and will not subsist.

3.1.6. Selection of Interesting Fixations

Once problematic fixations are identified as per the criteria above, these need to be removed from the EM dataset. The remaining fixations will become the time-locking events in the FRP data and will then be pre-processed according to the ET-guided EEG data and FRP data processing criteria.

3.2. PreProcessing of ET-Guided EEG Data

In the following, we present issues for consideration in respect of preprocessing of ET-guided EEG data regardless of experimental paradigm used and the type of processing (i.e., foveal or parafoveal) under investigation.

3.2.1 Data Synchronisation

In order to synchronise EM and EEG data and proceed with further data processing, the events that indicate trial onsets and trial offsets need to be coded with the same event markers both in the EM and EEG datasets (e.g., 31 for all TTLon/onset and 32 for all TTLoff/offset). If necessary, recoding may be required.

When both EM and EEG recordings include consistent event markers indicating the onset and offset of each trial, it is possible to synchronise the data. Synchronisation can be achieved via the EYE-EEG extension of EEGLAB (Dimigen et al., 2011), which also allows researchers to import the events present in the EM data (i.e., fixations, saccades, and blinks) as detected online by the ET system (at this point provided by SR Research EyeLink systems only) and add them to the EEG event structure. As we will discuss below, the imported EM events will be used to optimise the ICA procedure (i.e., for spike potential overweighting), to match the original timestamps of these events with the timing of the interesting fixation onsets, and to allow for the adoption of the deconvolution approach. It is important to note though that these EM events do not yet contain the features and selections identified at the end of the EM preprocessing procedures. That is, at this point all fixations, saccades, and blinks will be imported in the EEG data, regardless of whether they reflect interesting EM events associated with the experimental sentences, nor whether they are associated with the critical areas of interest in each experimental sentence.

Once EM and EEG data are synchronised, it is essential to verify that the two datasets perfectly match with each other. Perfect alignment is observed when the latency difference between the corresponding events of the two datasets has a mode of zero. Again, the EYE-EEG extension of EEGLAB (Dimigen et al., 2011) allows for checking and estimation of the synchronisation quality of the two recordings. In our experiments, we examine the synchronisation quality and error for each participant's dataset to ensure tight alignment.

Regardless of the paradigm used and of the type of processing investigated, synchronisation is crucial in co-registration research (see Baccino & Manunta, 2005; Degno et al., 2019a, 2019b; Dimigen et al., 2011; Kornrumpf et al., 2016; Kretzschmar et al., 2015; Loberg et al., 2018, 2019; López-Peréz et al., 2016; Metzner et al., 2015; Mirault et al., 2020; Niefind & Dimigen, 2016; Weiss et al., 2016 for details on data synchronisation).

3.2.2. Detection and Removal of Bad EEG Channels

During data acquisition, the signal quality for each channel should be maximised, and where possible, adjustments made to ensure high quality (e.g., by asking participants to adopt a comfortable position, thereby reducing movements, and in turn, muscle artifacts in the signal). However, when recording the EEG signal from multiple electrodes on the scalp surface or from special populations, it might happen that one or more channels may contain artefacts (detectable in the form of extreme values, e.g., a flat or a noisy channel), due, for example, to poor contact between the channel and the scalp. Such circumstances prevent the recording of a good signal from that channel, and this in turn affects the signal-to-noise ratio across the whole scalp. If a bad channel is present in the dataset, it is possible to remove it and at a later time, perform a (linear or spherical) interpolation, whereby the missing signal is estimated on the basis of the signals of the other (good) neighbouring electrodes (e.g., Perrin et al., 1989).

The detection of bad EEG channels can be performed in different ways, and in a manual or an automatized manner (e.g., autoreject, Jas et al., 2017; FASTER, Nolan et al., 2010; PREP, Bigdely-Shamlo et al., 2015). Commonly, EEG channels are considered bad when their values are particularly extreme and exceed a certain threshold (e.g., Junghöfer et al., 2000), or when the difference between minimum and maximum voltages at a particular electrode exceeds a certain threshold (i.e., peak-to-peak signal amplitude differences, e.g., Jas et al., 2017) (although see for example SNS, de Cheveigné & Simon, 2008, for a different method). The basic difference between the two being that the first approach estimates whether the values of each channel are within certain limits, while the second approach measures whether there is strong change for a particular electrode within each timeframe (e.g., epoch, sliding window, etc.), regardless of the overall voltage limits within the entire period of measurement. These two methods can be implemented in terms of absolute values (e.g., Junghöfer et al., 2000), or in terms of z scores (e.g., deviation and noisiness criteria, Bigdely-Shamlo et al., 2015; FASTER, Nolan et al., 2010). In addition, it is possible to detect bad EEG channels using a single 'global' threshold for all sensors and/or subjects, or to estimate multiple 'local' thresholds for each electrode and/or subject separately (see Jas et al., 2017 for a discussion).

We refrain from committing to a particular strategy to detect bad EEG channels, nor do we recommend particular threshold values as these might vary significantly between datasets, depending on the nature of the data at hand. However, it should be emphasised that if a global threshold is adopted, it is possible that one particular value might not be suitable to detect bad channels in all subjects. Thus, it may be necessary to spend time estimating the threshold that is optimal for the detection of bad EEG channels across all of the experimental datasets (assuming that most subjects' data will not contain bad sensors). Implementation of these methods can be found in popular software (e.g., EEGLAB, Delorme & Makeig, 2004; MNE-Python, Gramfort et al., 2013) and can be applied regardless of the experimental paradigm and type of processing under investigation in co-registration reading experiments (Degno et al., 2019a, 2019b; Loberg et al., 2018, 2019).

3.2.3. Initial Filtering

Filtering is used to improve the signal-to-noise ratio and increase statistical power for detection of true effects in the EEG data (e.g., Kappenman & Luck, 2010; see Luck, 2014 for a discussion of filters in relation to aliasing artefacts). By using filters, researchers can suppress signals in frequency ranges that are likely to be of no interest or may contain artifacts. For example, noise due to electrical power lines comprises a frequency of about 50 Hz for Europe and most of Asia, and a frequency of about 60 Hz for North America. Additionally, muscle contraction produces activity with frequencies predominantly above 100 Hz (although, depending on the type of muscle, the electromyogram spectrum can be much broader and include frequencies between

20-300 Hz; e.g., Muthukumaraswamy, 2013), and skin potentials (due, for example, to sweating) produce activity that consists mainly of frequencies lower than 0.1 Hz (e.g., Luck, 2014). Applying a filter to the EEG data can help to attenuate those frequencies with a non-neural origin, and therefore, noise.

There is an ongoing discussion in the literature about the use of filters (e.g., Acunzo et al., 2012; Rousselet, 2012; VanRullen, 2011; Widmann & Schröger, 2012). Therefore, it is far from our intention to give strict recommendations for filtering, on the assumption that filter parameters need to be carefully assessed according to the specifics of the experimental data and research question (see Widmann et al., 2015, for a discussion of filter design). Indeed, different co-registration reading studies have used different filter parameters (high-pass filter of 0.1 Hz Baccino & Manunta, 2005; Degno et al., 2019a; Degno et al., 2019b; Hutzler et al., 2007; Hutzler et al., 2013; López-Peréz et al., 2016; Mirault et al., 2020; Niefind & Dimigen, 2016; 0.2 Hz Dimigen et al., 2012; Kornrumpf et al., 2016; 0.25 Hz Dimigen et al., 2011; 0.3 Hz Kretzschmar et al., 2015; Metzner et al., 2015; Metzner et al., 2017; Simola et al., 2009; 0.5 Hz Henderson et al., 2013; Loberg et al., 2018, 2019; Weiss et al., 2016; low-pass filters of 20 Hz Kretzschmar et al., 2015; Loberg et al., 2018; 30 Hz Degno et al., 2019a, 2019b; Hutzler et al., 2013; López-Peréz et al., 2016; Loberg et al., 2019; 40 Hz Baccino & Manunta, 2005; Mirault et al., 2020; Niefind & Dimigen, 2016; Dimigen et al., 2012; Kornrumpf et al., 2016; Simola et al., 2009; 45 Hz Henderson et al., 2013; 70 Hz Dimigen et al., 2011; Hutzler et al., 2007; Metzner et al., 2017; Weiss et al., 2016; 100 Hz Metzner et al., 2015; and in some cases with a 50 Hz notch filter, Hutzler et al., 2007; Metzner et al., 2015; Metzner et al., 2017; Weiss et al., 2016).

However, we note that for typical cognitive experiments, a high-pass filter of 0.1 Hz is considered optimal in terms of the trade-off between benefits (i.e., signal-to-noise and statistical power improvement) and costs (i.e., attenuation of the signal and introduction of artefactual effects in the EEG waveform) (Kappenman & Luck, 2010; Kulke & Kulke, 2020; Tanner et al., 2015). If the recorded data is of sub-optimal quality (i.e., noisier data), a more stringent high-pass filtering level might need to be used (e.g., Maess et al., 2016).

In terms of low-pass filtering, at this point in the preprocessing pipeline, researchers might decide to apply a common 30 Hz low-pass filter to suppress frequencies typically associated with power line noise and muscle activity. However, as we will discuss below, to optimise ICA decomposition, keeping higher frequencies in the data (with a filter of at least 100 Hz) has proven to be effective (Dimigen, 2020). Should researchers wish to examine for example spectral phenomena, or should their data require different parameters, these filter levels should be optimized for the task at hand.

Finally, it is important to remind the reader that filtering creates edge artefacts at the beginning and end of the signal. To avoid edge artefacts occurring at the start and end of each epoch or average, filters can be applied to continuous EEG data (e.g., Widmann et al., 2015).

Researchers wishing to engage in co-registration reading experiments might be advised to consider this issue regardless of the paradigm and the nature of psychological processing that is investigated. Implementation of different types of filters is found in common EEG software (e.g., EEGLAB, Delorme & Makeig, 2004; Fieldtrip, Oostenveld et al., 2011; MNE-Python, Gramfort et al., 2013). For discussion of the different types of filters (e.g., finite vs. infinite impulse response, causal vs. non-causal filters) and their

shortcomings, we refer the reader to Luck (2014) and Rousselet (2012).

3.2.4. Independent Component Analysis (ICA)

EEG recordings during natural reading are contaminated by ocular artifacts (e.g., corneo-retinal dipole changes, saccadic spike potentials, eye lid artifacts, Plöchl et al., 2012). Yet, EMs are an integral part of co-registration reading experiments (indeed, in our view, eye movements are a central, integral, largely visually and cognitively determined, behavioural aspect of natural reading). Thus, distortion in the EEG signal due to EMs cannot all be discarded and must be corrected.

Several artefact correction methods have been proposed, including regression-based procedures (e.g., Croft & Barry, 2000a, 2000b; Elbert et al., 1985; Gratton et al., 1983), dipole modeling (e.g., Berg & Scherg, 1991; Lins et al., 1993), and blind source separation procedures (Iriarte et al., 2003; Joyce et al., 2004; ICA, Jung, Makeig, Humphries, et al., 2000; Kierkels et al., 2007; e.g., PCA, Lins et al., 1993). Of these techniques, ICA has become the most commonly used to correct for ocular artifacts in co-registration reading studies (e.g., Degno et al., 2019a; Degno et al., 2019b; Henderson et al., 2013; Hutzler et al., 2007; Hutzler et al., 2013; Loberg et al., 2018, 2019; Metzner et al., 2015; Metzner et al., 2017; Mirault et al., 2020; Weiss et al., 2016).

Independent component analysis (ICA) is a blind source separation method, which linearly decomposes the multi-channel time series data into maximally temporally independent signal components (e.g., Bell & Sejnowski, 1995). Each of these temporally independent components (ICs) acts as a spatial filter, such that it keeps the contribution of one physical source of the EEG signals, while filtering out the contributions of all the other sources (Makeig & Onton, 2009). The decomposition is data-driven, as this method does not require any *a priori* knowledge of the spatial and temporal properties of the ICs (Bell & Sejnowski, 1995). ICA is commonly used for removal of artefact signals, although it can be used to separate the brain sources that contribute to the scalp data as well (Jung, Makeig, Humphries, et al., 2000; Jung, Makeig, Westerfield, et al., 2000; Makeig et al., 1996; Onton & Makeig, 2006).

For a successful ICA decomposition, an adequate amount of data, both in terms of the number of samples and quality, needs to be computed. Below we discuss a series of factors that can affect ICA decomposition, and which might be considered in order to improve identification and isolation of EEG artifacts from neural activity patterns.

3.2.4.1. Detection of Bad EEG Intervals

Occasionally, EEG recordings might contain non-stereotypical voltage fluctuations, due, for example, to large body or cable movements. Inclusion of these bad intervals would inflate the dimensionality of the EEG data (jeopardising the ICA assumptions) and reduce the quality of the ICA output (e.g., Debener et al., 2010). In a manner similar to the detection of bad EEG channels, it is possible to detect bad EEG intervals using manual or automatized techniques that require absolute, or peak-to-peak, amplitude difference thresholds (e.g., autoreject, Jas et al., 2017; FASTER, Nolan et al., 2010; PREP, Bigdely-Shamlo et al., 2015). At this point in the preprocessing procedures, it is advisable to set up thresholds with high values to detect and remove these rare artifacts with unknown origin (Commonly Recorded Artifact Potentials, CRAP; Luck, 2014). In addition, intervals of continuous EEG data might correspond to

periods during which a blink occurred, or gaze was located at a point in space well beyond the spatial range of the presentation screen. As we explained in the EM preprocessing section, blinks are a source of noise which should be removed from the data. Likewise, when gaze is oriented to extreme positions, it is difficult to interpret such data in respect of the experimental task and cognitive processing. The EYE-EEG extension (Dimigen et al., 2011) can be used to detect these EEG intervals on the basis of the ET signal.

Finally, once bad EEG intervals are identified and marked down, it is possible to check the synchronisation lag between the ET and EEG systems with the cross-correlation procedure included in the EYE-EEG extension. This procedure makes use of the raw EEG and ET signal and computes the correlation between the raw horizontal ET channel and the horizontal EOG electrodes. The peak value of the cross-correlation is assumed to be the estimation of direction and amount of lag between the two systems. Thus, for an accurate synchronisation, the highest cross-correlation should be observed at a lag of zero samples.

3.2.4.2. Optimised ICA training data

It has been argued that the use of optimised training data helps with ICA decomposition (e.g., Debener et al., 2010; Dimigen, 2020). Training data are typically a copy of the original data on which different (stricter) parameters are applied. The ICA is then computed on the training data and the resulting component weights later applied to the original data (e.g., Meyberg et al., 2017). Dimigen (2020) has suggested that at least three parameters be considered for FRP training datasets: high-pass and low-pass filtering, as well as spike potential overweighting.

First, the contribution of low frequencies negatively affects ICA decomposition (Debener et al., 2010). These frequencies are associated with non-stationary signals that vary spatially over time, which compromises the ICA assumptions (for discussion about ICA assumptions see Makeig & Onton, 2009; Onton et al., 2006). Therefore, to optimise the ICA decomposition, a higher filter for EEG training data should be applied. Dimigen (2020) has evaluated different filter parameters and showed that in reading tasks, best results are obtained with the high-pass filters that are between passband edge of 2.5 Hz (width of transition band: 2 Hz, low cutoff (-6dB): 1.5 Hz) and 4 Hz (width of transition band: 2 Hz, low cutoff (-6dB): 3 Hz). In particular, best correction of the corneoretinal dipole artifact is obtained with a passband edge of 3-4 Hz, whilst best correction of the saccadic spike potential artifact is obtained when high-pass filter is about or less than 3 Hz (width of transition band: 2 Hz, low cutoff (-6dB): 2 Hz).

With respect to the low-pass filter, Dimigen (2020) has shown that when high frequencies (i.e., 100 Hz) are kept in the training data, the quality of ICA decomposition improves. The author suggested that the saccadic spike potential spectrum might therefore include high frequencies (> 40 Hz), and thus be better modelled when such frequencies are included in the training data (if source estimation analyses are not to be conducted).

Furthermore, it has been suggested that a factor that favourably affects ICA quality is the use of peri-saccadic samples overweighting (Keren et al., 2010). By appending segments of data extracted around saccade onset to the training data, the ICA can train on a larger amount of contaminated data, and, in turn, better separate between components of non-neural and neural origin. Dimigen (2020, see supplementary Figure S6) showed that in reading tasks cutting and mean-centering short 30 ms epochs (-20 to +10

ms from saccade onset), with the appended data length between 30-100% of the original training data length, almost fully removed the residual saccadic spike potential artefact from the data (see Dimigen, 2020 for details on the interactive effect of high-pass filtering and spike potential overweighting). The more training data that is used (and therefore the length of the appended data), the better the ICA decomposition. However, a decision is required regarding the percentage of data to use in order to balance computational power with better decomposition. When the number of data samples is not adequate, an alternative would be to compute first Principal Component Analysis (PCA) to reduce dimensions of the data (e.g., Kambhatla & Leen, 1997), and then perform the ICA decomposition, thereby obtaining a smaller number of ICs. However, recently, arguments have been put forward suggesting that using PCA for dimension reduction prior to ICA should be undertaken with caution and only after careful evaluation of whether it is necessary (Artoni et al., 2018).

The latest version of the EYE-EEG extension of EEGLAB (v0.85) includes functions to construct training data optimised for ICA decomposition.

3.2.4.3. ICA on Training Data

Running ICA requires a considerable amount of time. To speed up the process researchers might consider downsampling the training data. When downsampling the data though, consideration of the frequency content of the artefacts that are going to be modelled needs to be given. In other words, extreme downsampling, where the frequencies of the artefacts are no longer well represented (as defined through Nyquist frequency), should be avoided.

In addition, we remind the reader that the extended Infomax ICA algorithm (Bell & Sejnowski, 1995; Lee et al., 1999; Makeig et al., 1996) is the default ICA algorithm used in EEGLAB (Delorme & Makeig, 2004). However, there are multiple options available for ICA decomposition (e.g., FastICA, Hyvärinen, 1999; JADE, Cardoso & Souloumiac, 1993). We refer the reader to the original papers for details of the different algorithms.

3.2.4.4. Detection of Ocular ICs on Training Data

The identification of ICs associated with ocular artefacts has been a matter of long debate, and without the use of the computational techniques we will describe below, it can be considered quite a subjective practice.

The EYE-EEG extension (Dimigen et al., 2011) that has been implemented in EEGLAB (Delorme & Makeig, 2004) allows researchers to identify the ICs in a more objective way. This approach implements the variance-ratio criterion proposed by Plöchl et al. (2012), such that ICs likely to reflect ocular artefacts show a higher temporal variance-ratio during a saccade than during a fixation. Dimigen (2020) has tested different variance ratio thresholds (between 0.6 and 1.5) and showed that a threshold no lower than 1.1 should be adopted to adequately remove ocular ICs without removing neural activity relevant to the cognitive task at hand (i.e., overcorrection).

3.2.4.5. ICA Weights Transfer

Once the ICA weights are computed and ocular ICs are identified in the optimised training dataset, the weight matrix is applied to the original EEG data. This procedure allows the researcher to strike a balance between the requirements of an effective ICA decomposition (i.e., with strict high-pass filtering) and the requirements of observing the

non-artifactual brain activity in the low frequencies of the data (i.e., with lax high-pass filtering).

3.2.5. Pruning of ocular ICs

Once the original EEG data contain the ICA weights and the identified ocular ICs, it is straightforward to remove the ocular ICs, and plot the data with and without those ICs (see the EYE-EEG tutorial "Step 8: Remove ocular artifacts with eye tracker-guided, optimized ICA" at http://www2.hu-berlin.de/eyetracking-eeg/tutorial.html#tutorial6).

3.2.6. Low-Pass Filtering

Low-pass filtering is used to attenuate higher frequencies, for example associated with line noise (i.e., 50 Hz for Europe and most of Asia, and 60 Hz for North America) and muscle artifacts (typically >100 Hz). After identification and pruning of ocular ICs, which required higher frequencies to remain in the data, researchers might want to apply a stricter low-pass filter. A low-pass filter of 30 Hz is commonly used in typical cognitive experiments (e.g., see Figure A5 of Rousselet, 2012, for frequencies of filter cutoffs) to increase the signal-to-noise ratio and filter out high frequencies associated with muscle artifacts, as well as the 50 Hz (or 60 Hz) power line noise, so that no separate notch filter for electric noise needs to be implemented.

However, as for most of the decisions presented in the pipeline, the parameters to be adopted will vary depending on the research questions to be investigated (e.g., examination of component onset latencies).

3.2.7. Interpolation of Removed Bad EEG Channels

We discussed at the beginning of the present pipeline that at times it is necessary to remove a few bad EEG channels in some subjects, as otherwise the noisy data in these channels might negatively affect further steps in the data preprocessing (e.g., ICA, re-referencing, etc.). It is likely that the bad channels that have been removed will be different for different subjects. Thus, unless a single sensor is to be analysed, and for none of the subjects that sensor is a bad channel, it is necessary to restore data dimensionality across subjects. To achieve this, it is possible to estimate the missing data of the bad EEG channels using either a linear or spherical interpolation (Perrin et al., 1987, 1989).

Interpolation can be performed in all the popular software for EEG analysis (e.g., EEGLAB, Delorme & Makeig, 2004; Fieldtrip, Oostenveld et al., 2011; MNE-Python, Gramfort et al., 2013).

3.2.8. Detection of Bad EEG Intervals

Intervals that still contain extreme values after the previous preprocessing steps are identified at this point in the pipeline through the adoption of stricter criteria. As for the initial detection of bad EEG intervals, extreme values could be detected in absolute value or as a difference between a minimum and maximum voltage at each electrode, using the same threshold for all subjects, or different thresholds for each subject.

3.2.9. Re-Referencing

The voltage that we record at each scalp electrode is measured with respect to a reference site (i.e., [Active electrode – Ground electrode] – [Reference electrode – Ground

electrode]), which is assumed to be electrically neutral (i.e., unaffected by the sources of interest and noise). However, none of the sites typically used as a reference site (e.g., vertex, tip of the nose, earlobes or mastoids) is completely neutral. All of them are still affected by some artefacts and/or brain signal to some degree. Thus, offline re-referencing procedures are generally used to minimise the impact of the activity at the reference site upon the EEG signal recorded at different scalp electrodes.

The choice of which reference to use is an issue of discussion in the EEG literature (e.g., Kayser & Tenke, 2010), and it depends on one's experimental procedures and research questions. The most commonly used offline procedures for co-registration reading experiment data analysis are linked mastoid (Kretzschmar et al., 2009; Kretzschmar et al., 2015; López-Peréz et al., 2016; Metzner et al., 2017; Simola et al., 2009) and average (Degno et al., 2019a, 2019b; Dimigen et al., 2011; Dimigen et al., 2012; Kornrumpf et al., 2016; Loberg et al., 2018, 2019; Metzner et al., 2015; Niefind & Dimigen, 2016) re-referencing procedures.

The linked mastoid re-referencing procedure is computed as the average between the left and right mastoids (e.g., Mahajan et al., 2017). It is commonly adopted in the ERP/FRP community as it is independent of electrode montages (i.e., facilitating comparison between different laboratory set ups), and because the location of the mastoids is considered to be sufficiently distant from the neural sources of the EEG signal. However, Li et al. (2015) showed that when cortical sources generating an effect are located in the inferior occipito-temporal regions (i.e., the cortical sources of the preview positivity effect as estimated by Dimigen et al., 2012), the use of linked mastoids as offline re-referencing procedure can attenuate the investigated effect. Under these circumstances, an average re-referencing procedure is a better option.

The average re-referencing procedure (Bertrand et al., 1985) is computed as the average of the activity across all the scalp electrodes. This approach is considered good if enough electrodes are used on most of the accessible parts of the head (e.g., at least 60 sites across the entire scalp; Dien, 1998, Picton et al., 2000; although see also Desmedt et al., 1990 for a related discussion of pitfalls of average referencing).

Offline re-referencing is a straightforward procedure, and its implementation can be found in all the popular software (e.g., EEGLAB, Delorme & Makeig, 2004; Fieldtrip, Oostenveld et al., 2011; MNE-Python, Gramfort et al., 2013).

3.3. Processing of FRP Data

3.3.1. Matching the EM and EEG Data

In the present pipeline, at this point we identify in the EEG data the same interesting fixations that were selected at the end of the EM data preprocessing. This is a critical step as it allows for the use of these interesting fixations as time-locking events for the FRP data. Similarly, if researchers are interested in saccadic-related potentials (SRPs) rather than FRPs, and the critical saccades have been identified in the EM data preprocessing, it will be possible at this point to select the saccade onsets of interest to time-lock the SRP data.

3.3.2 Fixation-Related Potential Estimation

In natural reading experiments, words are fixated in

rapid succession, with very short intervals between one fix-ation and the following one, and participants are free to move their eyes in any direction and location. Under these circumstances, the FRP waveform elicited by the processing of each word will temporally overlap with the FRP wave-forms associated with the processing of previous and fol-lowing words. In addition, different fixation durations and saccade sizes will be present. When overlapping potentials and oculomotor behavior systematically differ between conditions, these differences can sometimes result in spu-rious effects in the FRP data, particularly in respect of later components where EM behaviour may differ quite markedly contingent on the experimental condition within which it was acquired (e.g., Dimigen & Ehinger, 2020). It is very dif-ficult, if not impossible, to avoid the overlapping issue and control for oculomotor confounds during data acquisition in natural reading tasks. However, some techniques have been proposed to deal with these two issues during offline data preprocessing.

Nikolaev and colleagues (2016) proposed to match the EM characteristics between experimental conditions based on the Mahalanobis distance (i.e., a metric of the distance between a data point and the mean of the multivariate dis-tribution with a covariate matrix; Mahalanobis, 1936) or ranked-based Mahalanobis distance when the data have a highly skewed distribution (i.e., the distance between two data points with an adjusted covariate distribution; Rosen-baum, 2005) in order to counterbalance these oculomotor differences. That is, first, to set minimum EM criteria (e.g., minimum fixation duration of 250 ms), and then identify a subset of the data and use only those observations in which EM characteristics (e.g., fixation duration, saccade amplitude and saccade direction) are not significantly dif-ferent across conditions, as interesting time-locking events (i.e., interesting fixations or saccades) for the FRPs. In the event that a different number of observations might survive across different conditions, observations might be random-ly selected from the condition with the highest number of observations, such that the data set(s) match the condition with the lowest amount of trials.

Although the proposed solution may attenuate the ocu-lomotor behaviour variability issue, it has the obvious weakness of requiring that a population of short fixations that could reflect meaningful cognitive processes will be ex-cluded from the analyses. Also, by including exclusively ob-servations with first fixation durations longer than a mini-mum time window (e.g., 250 ms), the variability issue would still hold true for late FRP components, wherein different fixation durations, regressions and refixations may lead to differently overlapping potentials. An alternative action of selecting observations exclusively for very long fixation du-rations whilst taking into account overlapping late FRP components, would limit the analysis to a very small, and likely non-representative, sample of the data (given that on average an eye fixation is approximately 250 ms). Further-more, and very importantly, differences in fixation duration or saccade size may be fundamentally associated with the different experimental conditions (e.g., processing in one condition might be more difficult than in another resulting in longer fixation durations). Finally, until we have a clear and detailed understanding of the relationship between EM

and FRP effects, some ambiguity will remain with respect to the consequences that EM matching procedures might have on the FRP data. Thus, reducing the data to those observa-tions that have a similar fixation duration or saccade size across conditions may actually bias the results attenuating and/or shifting the onset of any real FRP difference[1].

Regression-based approaches have also been recom-mended to deconvolve overlapping ERP/FRP signals and, at the same time, to model influences of variables that cannot be fully controlled experimentally (e.g., Cornelissen et al., 2019; Kristensen et al., 2017; Smith & Kutas, 2015a, 2015b; c.f. Frey et al., 2013 for adjacent response estimation algo-rithm, ADJAR; Woldorff 1993, to estimate FRPs). Recently, a Matlab toolbox has also been developed to perform such analyses (Unfold; Ehinger & Dimigen, 2019; although see also the existing mTRF, Crosse et al., 2016, and MNE, Gram-fort et al., 2013).

The Unfold toolbox (Ehinger & Dimigen, 2019) is written in Matlab and combines general linear models to estimate ERP/FRP response and deconvolve overlapping signals, and generalised additive models to model non-linear predictors. We refer the reader to the original paper for an extensive description of assumptions, mathematical details, and fea-tures available in the toolbox. Here we limit ourselves to note two major implications for data preprocessing that re-searchers need to know when deciding whether to use this toolbox.

First, deconvolution is implemented in the Unfold tool-box such that it requires FRP data to be continuous. This implies that researchers will need to use methods able to deal with artefact correction in continuous rather than epoched data. For example, when trials include intervals with bad EEG data, those intervals (or trials) will need to be identified and marked down, but not removed from the con-tinuous dataset. Removal of the bad EEG intervals (or trials) from the dataset would lead to a less reliable model fit of the overlapping potentials, as the associated events would also be removed, and therefore not accounted for. Instead, if bad data are present in the dataset, they will be kept but marked down by setting the parameters of the equation (or time ex-panded matrix) explaining those samples to zero (i.e., so that those samples cannot contribute to FRP estimation).

Second, when using regression-based approaches such as in Unfold, variables of interest and covariates do not need to be categorical, as in the traditional average ERP/ FRP studies. Predictors can be either continuous or cate-gorical, making experimental design more flexible (Smith & Kutas, 2015a).

To date, a very limited number of co-registration reading studies have been published using regression-based ap-proaches to deal with overlapping signals and differences in stimuli characteristics (i.e., word length, Loberg et al., 2019) or oculomotor behaviour (i.e., saccade amplitude and fixation duration, Weiss et al., 2016) and to show how real and spurious effects can be detangled (Dimigen & Ehinger, 2020). Thus, systematic investigation is required in the fu-ture to clarify potential impact that differences between conditions in oculomotor behaviour and signal overlap have on our interpretations and conclusions about the investi-gated cognitive processing.

To illustrate the steps that may be followed in the

---

1 We note, however, that when separate analyses were carried out to investigate whether same or different effects were observed in the ear-ly FRP components associated with observations longer than a minimum duration (i.e., 240 ms), or with all the observations regardless of first fixation duration, the same effects were found (Dimigen et al., 2012).

pipeline when adopting, or not adopting, the deconvolution approach, we have provided two separate workflows (see Figure 1). The construction of the pipeline allows the decision between the traditional averaging approach and the deconvolution approach to be left as late as possible.

We note that the Unfold toolbox provides the option of running the same model with and without correction for overlapping signals. Without overlap correction the procedure is the standard rERP estimation procedure as described in Smith and Kutas (2015a), allowing for control of confounding variables but not for signal overlap. Thus, comparing FRPs with and without overlap correction can be a useful means of examining the consequences of overlap. Below we will discuss early considerations that may be useful when adopting deconvolution approaches, and then the steps to consider when adopting the more traditional ERP/FRP procedures. Baseline correction applies similarly to both approaches, for this reason we will discuss this issue and step in the same section.

3.3.3. Deconvolved Fixation-Related Potential Estimates

To estimate the subject-level FRP waveforms (*betas*), it is necessary to define the model that will best account for overlapping responses and best describe the effects of interest in the data. Dealing first with the overlapping issue, the deconvolution approach permits a regression equation with a time expanded matrix for each sample of the dataset to be built. The matrix is defined by the events and their timing in the data (e.g., interesting and non-interesting fixations) and is expanded to include a window of time locked to each interesting and non-interesting event where overlapping signal is estimated to occur (and to include the latency of the FRP effect to be investigated). By including a time window long enough to include neighbouring events, the model can reconstruct the responses that generated the whole dataset while accounting for the overlap.

Second, researchers will need to define a model that includes all the variables of interest and potentially confounding covariates, specifying whether these predictors are categorical/factorial or continuous variables, and if continuous, whether there is a linear or non-linear relationship between each predictor and the FRP response. In addition, the model will need to include the type of effects to be investigated (i.e., main effects, interactions, or both).

Once the model is defined and solved, a *beta*, that is a regression coefficient, for each parameter and timepoint of the defined time window will be returned. With the Unfold toolbox (Ehinger & Dimigen, 2019) the models are defined with the Wilkinson notation, spline regressions are used to define non-linear predictors, and *betas* can be exported in different formats (e.g., plain text, Matlab file, Fieldtrip structure) to be plotted and used in further statistical analyses. These features make Unfold an accessible tool for researchers to run these types of analyses.

3.3.4. Baseline Correction (with and without Deconvolution)

Baseline correction involves subtracting the average EEG activity recorded in the baseline time period from each channel and time point of the entire epoch (e.g., Urbach & Kutas, 2006). The purpose of the baseline correction is to reduce offsets and drifts and measure real changes in the EEG activity following the event of interest (e.g., word onset or fixation onset). Thus, for the baseline correction to be effective, the assumption is that activation during the baseline period is unaffected by the experimental manipulation and does not differ between conditions in any way (e.g., in terms of external environment). Indeed, any pre-event difference that might exist would lead to a false difference between conditions during the post-event interval, which could lead to the misinterpretation of changes in EEG activity as causal consequences of the event of interest.

According to Luck (2014) a good baseline period should be at least 20% of the overall epoch duration and a multiple of 100 ms in order to counteract the EEG oscillations due to alpha-frequency (~10 Hz). Nonetheless, different common practices are used. In ERP experiments, it is common to consider as baseline the time-window immediately preceding the stimulus presentation, usually the preceding 100 ms (e.g. Dambacher et al., 2006; Laszlo & Federmeier, 2009). This 'individual baseline' (Nikolaev et al., 2016) might represent a good choice for ERP experiments, as in these studies the inter-stimulus onset is typically quite long, and thus we can assume that the activity pre-stimulus is very similar across conditions.

For co-registration experiments, the choice of baseline is a little more complicated. To date many co-registration reading studies have used the same baseline period as that used in traditional ERP studies (e.g. Baccino & Manunta, 2005; Degno et al., 2019b; Dimigen et al., 2011, 2012; Simola et al., 2009). However, assuming preprocessing of the word to the right of the fixation (i.e., parafoveal processing of a target word; see Rayner, 1998, 2009 for reviews), the cognitive processing that occurs in the time window prior to the fixation onset on the target word would (potentially) be different across conditions due to the experimental manipulation of the target word. The extent to which this occurs would depend on the type of information that was available and extracted from the parafovea. In addition, systematic differences between conditions due to the overlapping issue discussed above would still apply.

With the regression-based approaches that include deconvolution, the choice of baseline might be less problematic. The deconvolution should remove the systematic variations between conditions due to overlap in the baseline interval, making it possible to use the interval preceding the fixation onset on the target word. Nonetheless, it is important to consider how baseline selection might be optimised with respect to the specific characteristics of a particular experimental paradigm to be employed to investigate specific research questions. When investigating parafoveal processing for example, it might be sensible to select a different baseline period to avoid any difference due to the processing of the target word which may have already commenced. Such a period might be the time window preceding the fixation onset on the pre-target word. That is, the fixation before the word that appeared prior to the target in the sentence, as up until this point the content of the sentence across conditions was almost always identical (Degno et al., 2019a; see also Coco et al., 2020 for a similar approach in visual search). Alternatively, for deconvolved FRPs, one could use a period of time preceding the fixation on the target word prior to it being processed (e.g., from -700 to -500 ms from fixation onset on the target word; Loberg et al., 2019). This approach, also known as 'local baseline' (Nikolaev et al., 2016), is ideal when the investigation focuses on processing of one or two specific target words within a sentence. However, often paradigms are employed in which every word within a trial (e.g., a list of words or a sentence) is manipulated in some way. Although current practice in these situations is to adopt a baseline of 100 ms from the

fixation onset of each word, this approach can be argued to be susceptible to the shortcomings discussed above. It remains an issue for future research to establish the best form of baseline (and the possible implications of current practice) in relation to future co-registration studies of reading.

A non-standard procedure of computing an 'individual baseline' involves taking a time interval immediately following stimulus presentation or fixation onset, generally the first 20 or 100 ms (e.g., Hutzler et al., 2007; Rämä & Baccino, 2010) as the baseline period. However, this approach is only valid when it is known that the effect of interest does not occur during that baseline period. Otherwise, this approach too would be susceptible to the shortcomings identified and might only be advisable when systematic differences between conditions and signal overlapping are accounted for. An alternative perspective is that baseline correction is not always necessary and, if performed, it should be included as a covariate in the statistical analysis to determine the strength of the correction (Alday, 2019).

To conclude, the choice of an appropriate baseline requires very careful consideration in relation to the experimental design that is employed, and with respect to whether the effect of interest might be expected to persist over several fixations. Baseline correction is implemented in all the commercially available and open source software for EEG data preprocessing (e.g., EEGLAB, Delorme & Makeig, 2004; Fieldtrip, Oostenveld et al., 2011; MNE-Python, Gramfort et al., 2013), although with varying flexibility.

3.3.5. Epochs (without Deconvolution)

Epoching involves extracting time windows that are time-locked to the event of interest (e.g., first fixation onset on the target words). Each epoch has a fixed length that includes a period of time before and after the event occurs. When extracting the epochs, one needs to consider which baseline will be used, and which FRP components will be investigated, as these latency periods will need to be included in the epoch length. It is also important to remember that at this point in the preprocessing, it can be helpful to extract longer epochs that extend beyond the end of the period to be examined, as these may be shortened at a later time.

At this point in the pipeline, researchers should ensure that the signal does not contain any bad intervals. If extreme values are present, additional adjustments should be made (e.g., rejection of intervals with extreme values as discussed above). Furthermore, if baseline correction is applied, this is typically done after epoching the data.

3.3.6. Averages (without Deconvolution)

Once the data have been time-locked to the events of interest (e.g., first fixation onset) and epoched, it is possible and straightforward to average the signal. Averaging means summing together the single-trial EEG waveforms at each timeframe and dividing by the number of epochs (e.g., Luck, 2014). Different types of averaging can be performed (e.g., averaging the EEG waveforms of all epochs belonging to one condition across time points within a time window, across electrode sites located on a scalp region of interest, etc.). Each condition will then be entered in the statistical analysis as a level of the predictor.

4. Statistical Analysis

In order to conduct statistical analyses on the final EM and FRP data, it is necessary to separate the two data sets and treat them independently. This will allow researchers to examine the investigated effects in both the EM and FRP data. We note that with deconvolution the number of interesting fixations that are extracted might be higher than the number of interesting fixations that are extracted with the non-deconvolution approach. If bad intervals of FRP data are present, the deconvolution model would simply ignore that bad segment for FRP estimation but still use the remaining intervals to model the effects of interest and the overlap. Thus, unless the fixation onset event overlaps with the bad data interval, that event will survive and enter the EM statistical analyses. In these instances, however, EM data should not be distorted, as the EM information (e.g., first fixation duration, gaze duration, etc.) will remain unmodified. In contrast, when adopting the non-deconvolution approach, if bad FRP data are present, the entire FRP epoch is removed from the FRP dataset, and therefore, the interesting fixation associated with that epoch will not survive and will not contribute to the EM statistical analyses.

In this section we discuss two statistical methods that researchers might consider using when dealing with co-registration data, that is cluster-based permutation tests for the FRP data, and linear-mixed effects models for both EM and FRP data. We acknowledge that the analysis of variance (ANOVA) is probably the most common statistical method used in psychological sciences, with respect to both the early studies recording EM measures (e.g., Rayner, 1977), and the early and more recent experiments recording ERP/FRP measures (e.g., Baccino & Manunta, 2005; Kretzschmar et al., 2015; Kutas & Hillyard, 1980). However, in recent years, researchers in both EM and ERP fields are adopting alternative statistical approaches (e.g., Baayen et al., 2008; Ehinger & Dimigen, 2019; Kliegl et al., 2011; Smith & Kutas, 2015a, 2015b). Therefore, quite deliberately, we will not cover ANOVA in this paper. Instead, we will focus on the statistical approaches that we have been guided towards via the review process with respect to work we have published. We will deal with specific issues associated with these approaches below. Depending on the type of statistical test one is planning to conduct, the structure of the data to be extracted will differ. We will discuss this aspect of data preparation according to each statistical approach. We note that these two statistical methods do not address the problem of overlapping potentials. However, they can both be combined with deconvolution models (although less easily for linear-mixed effects models; Ehinger & Dimigen, 2019; Ehinger, 2019).

4.1. Cluster-Based Permutation Statistics

This approach is a form of non-parametric mass univariate method in which a t-test is performed between different conditions at each sample, that is, at each electrode site and at each time point, whilst controlling for multiple comparisons (Groppe et al., 2011; Maris & Oostenveld, 2007). In this approach the basic ideas are that (1) if observations from different conditions are drawn from the same probability distribution, then the different observations are interchangeable, (2) if the EEG signal differs between experimental conditions at one electrode and at one point in time, it is very likely that adjacent electrodes and time points will also exhibit a similar effect. With respect to the first assumption, as in all non-parametric permutation tests, per-

muting the data involves assigning the actual observations randomly to conditions, after which the test statistic is calculated on this random data set. This process is performed a large number of times (typically between 1,000 and 10,000 times), in order to obtain an accurate null distribution of the permuted data (i.e., the permutation distribution). The observed test statistic is then compared with the random test statistics drawn from the permutation distribution, and if the observed test statistic is larger than the random test statistics, the experimental hypothesis is accepted. Regarding the second assumption, that is, similar effects will occur over adjacent time-electrode samples, spatio-temporal clusters are identified. Thus, t-values are computed for each (electrode-time) sample, and those samples with t-values larger than a threshold (e.g., values associated with p < .05) which exhibit a similar effect (e.g., with the same negative or positive sign) are clustered according to their temporal and spatial adjacency. The sum of the t-values within each cluster forms each of the cluster-level statistics. The largest absolute value amongst all of the cluster-level statistics is then evaluated against the values within the permutation distribution, to establish where within that distribution it falls. Again, if the observed cluster-level statistic is greater than a threshold (e.g., the 95% point) in the cluster-level permuted distribution, then the experimental hypothesis is accepted.

The advantages of this approach are that, first, it can handle non-normal data, in fact no assumption is made on the parameter estimates. On the contrary, it uses the properties of the actual observed data for the evaluation of statistical significance. Second, it is data-driven, as all electrode sites and time points can be entered into the analysis, and it is not necessary to pre-select specific scalp locations and time windows. Thus, researchers might consider adopting this approach when a study is sufficiently novel or exploratory in that there are no *a priori* predictions concerning the expected latency of effects (i.e., the temporal windows for analyses) and expected locations at which the effects should occur (i.e., electrode sites). Note, though, that beyond this, the technique may be employed in more standard experimental situations, where researchers could limit their analyses to specific time windows and small numbers of electrode sites on the basis of prior expectations (although see for example Maris & Oostenveld, 2007; Sassenhagen & Draschkow, 2019 for a discussion of inferential output misinterpretation). This will reduce the number of statistical tests (and thereby, the number of corrections for multiple comparisons), and increase, in turn, the sensitivity of the test. In this situation, the time windows and electrodes to be included in the analyses should be made in the absence of knowledge regarding the data, and a clear, *a priori* justification of the choice should be provided (e.g., Luck & Gaspelin, 2017). Third, cluster-based permutation tests provide a strong degree of certainty in detecting the presence of an effect (i.e., by controlling Type I error rate; Pernet et al., 2015). Arguably however, a disadvantage of this approach is that, given the large number of comparisons, then the corrections that are applied mean that the test is quite conservative, and some small or short-lived effects may be missed.

Given the nature of this statistical analysis, the approach works well with high density EEG caps (i.e., recording from >60 electrodes). If recording with less channels, one could still perform cluster-based permutation statistics, but it may be better to cluster over time only, rather than over both space and time (or with a different cluster neighbourhood construction).

Software with relatively straightforward implementation and well-documented tutorials to conduct cluster-based permutation tests is Fieldtrip (Oostenveld et al., 2011; although see for example also MNE-Python, Gramfort et al., 2013). To extract FRP data to be entered in Fieldtrip and run this type of analysis, it is sufficient to calculate condition averages for each subject (i.e., to be used for statistical tests) and then condition grand averages across subjects (i.e., to be used for plotting the data). The data can be converted back and forth from EEGLAB to Fieldtrip readily with specific functions (e.g., *fieldtrip2eeglab* and *eeglab2fieldtrip*).

4.2. Linear Mixed Effects Models

This approach is a multilevel regression model in which both fixed (e.g., the independent variables) and multiple random (e.g., participants and items) factors are incorporated (Hox, 2010; Pinherio & Bates, 2004). The main advantages of this method are that (i) it can handle unbalanced designs, as it weights the contribution of each participant, (ii) it incorporates multiple random factors in one single model, (iii) it can easily include and deal with multiple covariates, and (iv) the independent variables can be either categorical or continuous factors.

Linear mixed effects models are currently the most common type of statistical analysis used in EM research. In fact, data loss is a situation that commonly arises in EM data, due for example to a different number of blinks, skips, or bad trials between conditions. Typically, separate models are computed to analyse different EM measures (e.g., first fixation duration, gaze duration, etc.; although see von der Malsburg & Angele, 2017 for an alternative perspective on the practice of testing multiple dependent EM measures). This is ordinarily done with a quite specific purpose in mind, namely, to examine the time course of experimental effects across EMs as they are made through the sentence and over time. To this extent, multiple EM measures are very valuable to researchers (and together different measures can inform theoretical understanding to a greater extent than when each is considered in isolation).

This approach has great potential to also become a common statistical tool to analyse electrophysiological data (e.g., Baayen et al., 2008; Bagiella et al., 2000). However, to date, few studies have analysed ERP or FRP measures using this approach (e.g., see Amsel, 2011; Dimigen et al., 2011; Kornrumpf et al., 2016 for visual word recognition studies with ERP/FRP; see Fröber et al., 2017; Frömer et al., 2016a, 2016b for studies recording ERPs in other domains). Linear mixed effects models are not run on a table of means, but each trial for each participant is entered into the analysis as a single data point (i.e., single-trial ERP analyses). Given that EEG data take the form of a stream of voltages over time and at multiple channels, potentially, the amount of ERP/FRP data to be entered into a linear mixed model can be substantive, and this in turn could make this form of analysis impractical. The impracticalities would arise because of the volume of data and the necessary computational power to handle it, as well as the possibility that Type I (false positives) errors would arise.

An alternative approach would be to average the ERP/FRP signals across time windows and/or electrode clusters that are defined a priori (e.g., with registered reports; Chambers, 2013), to reduce the complexity of the results (e.g., Niefind & Dimigen, 2016). However, the drawback here would be that researchers would need to be quite confident about the measurement parameters (i.e., electrode

sites and latency windows) (Kriegeskorte et al., 2009). Furthermore, since data in this type of analysis are necessarily averaged, the level of detail of the spatial and temporal dimensions of the effects would be reduced. A third scenario would be to run separate linear mixed effects models for a few specific electrode sites and/or time windows (e.g., Kornrumpf et al., 2016). This approach would avoid the need to run a large number of separate models, and at the same time would provide a degree of detail regarding the temporal and spatial dimensions of the effects. However, once again, with this approach the measurement parameters need to be known in advance, and the ERP/FRP measures may not provide a comprehensive picture of effects, as a significant proportion of the temporal window and electrodes would not be considered in the analyses. Therefore, perhaps this approach may be considered most appropriate when there are several existing studies that are quite similar in nature in the literature that provide a point of reference for measurement parameters.

A recent suggestion has been to combine cluster-based permutation tests with linear mixed effects models (Frömer et al., 2018). The idea is to use the results from the cluster-based permutation analysis as a guideline for the choice of the electrode sites and time windows that should be entered into the ERP/FRP analyses with linear mixed effects models. This approach could overcome some of the current limitations of each of these statistical tools alone. However, as pointed out by Kriegeskorte et al. (2009), this itself could introduce the issue of circular analysis and 'double dipping', in addition to the 'inapplicable' use of the cluster-based permutation tests to infer a very precise spatio-temporal extent of the effects (e.g., Sassenhagen & Draschkow, 2019). At this stage, more research needs to be conducted in order to establish whether, and how, Frömer at al.'s suggestion offers a viable possibility into the future, and to identify alternative tools for analyses of these complex co-registration data sets.

A variety of software packages to perform mixed effects models is now widely available, and among these the open source statistical programming environment R (Baayen et al., 2008; Ihaka & Gentleman, 1996) is popular, being very flexible and for which good documentation is available. To extract EM and FRP data for analysis with linear mixed models in R, it is necessary to export all the variables which will be used for statistical analysis. For the FRP data in particular, this implies that it is essential to decide whether mean amplitudes or peaks, or both, will be analysed. Second, it is important to decide whether clusters of electrodes or single electrodes, and time windows or single time points are to be considered for analyses. For example, researchers could identify clusters of electrodes and time windows where the effects are expected, and average across electrodes and time windows (Niefind & Dimigen, 2016). Alternatively, researchers could extract the FRP data from single electrodes and average across time windows (Kornrumpf et al., 2016), or else data from single electrodes and specific time points could be extracted. The choice of the type of data to be extracted will depend on the research question and whether there are existing experimental findings on which to generate *a priori* predictions for specific types of effects at specific scalp locations and points in time.

5. Implementation

One of the aims of the present paper was to provide researchers who are wishing to engage in co-registration reading research with some scripts they could use as the ba-

sis for data analysis. Thus, after discussing the steps of the pipeline that require consideration, as an example, here we present the decision steps that we took in relation to a sample dataset from one of our own studies, and we present how we have implemented each of these decisions. Again, these decisions were made in relation to our specific data, and may not be appropriate choices for all datasets, paradigms and research questions. Parameters should be modified to reflect particular data sets and research questions. By giving details of our decisions and scripts, and by allowing the reader to check the lines of code corresponding to our decisions, we hope that this will act as an example and a guide through the complexity of co-registration data analyses.

5.1 Sample Datasets

We provide six co-registration datasets (i.e., EM and EEG data) from a sentence reading experiment by Degno et al. (2019a). The datasets should allow the reader to test each step of the pipeline. We note however that the pipeline we present here is slightly different to that used in Degno et al. (2019a), and this deviation reflects continued improvement in our approach in line with up-to-date procedures and discussion within the scientific community.

The data were recorded in a sentence reading experiment. Single-line sentences were presented on a computer display and participants were required to silently read each sentence for comprehension. Two target words in each sentence were manipulated using the boundary paradigm (Rayner, 1975), whereby parafoveal preview of each of the target words was experimentally controlled. Prior to direct fixation, the preview of each target word was a string of Xs, a string of random letters, or an identity preview that was identical to the target word. Upon direct fixation, the target words could be high or low frequency words. The experiment aimed to investigate parafoveal and foveal effects of target preview mask and target word frequency.

Connection between the ET and EEG systems was made through a parallel port connecting the stimulus presentation computer running Experiment Builder (SR Research) and the EEG amplifier (Neuroscan Compumedics). Synchronisation of the two data streams was achieved by presenting a TTL pulse sent from the stimuli display computer running Experiment Builder (SR Research) at both the beginning and at the end of each trial. The TTL pulse was recorded as a message in the EM data (e.g., '*TTLon112*' and '*TTLoff112*' for the onset and offset of trial index 112, in the column "CURRENT_MSG_TEXT" of the SR Research DataViewer message report) and as trigger in the EEG data (e.g., 3 for the onset and 103 for the offset of a trial in the experimental condition 3 in the EEG event file).

Monocular EM data were recorded with an Eyelink 1000 tracker (SR Research) with a sampling rate of 1000 Hz. The EEG data were recorded with Scan 4.5 (Neuroscan Compumedics) from 64 scalp electrodes and 4 bipolar EOG channels located according to the 10-20 SI, with SynAmpsRT amplifiers (Compumedics Neuroscan) in a DC mode at 1000 Hz and low pass filtered online at 100 Hz, and with the tip of the nose as online reference. The display computer was a 19-in. CRT screen, with a resolution of 1,024 x 768 and a refresh rate of 140 Hz.

5.2 Scripts

The script we provide for the preprocessing of the EM data was created and run in R. The input files include fixation,

saccade and message reports (extracted from Data Viewer, SR Research), and the EEG event files (extracted from Neuroscan). The EM parsing was automatically performed by the Eyelink 1000 tracker (SR Research).

The script we provide for the preprocessing of the ET-guided EEG data preprocessing and FRP data preprocessing was created and run in Matlab (versions 2018a update 6 and 2018b), and assumes the presence of the EEGLAB toolbox (version v14.1.2; Delorme & Makeig, 2004), the EYE-EEG extension (version 0.85, http://www2.hu-berlin.de/eye-tracking-eeg; Dimigen et al., 2011), the Unfold toolbox (version 1.1, Ehinger & Dimigen, 2019), and the Fieldtrip toolbox (Oostenveld et al., 2011) in the Matlab path. The input files include the raw EM data in ASCII format (extracted from Data Viewer, SR Research), the raw EEG data (extracted from Scan 4.5, Neuroscan Compumedics), and a channel location file. In the FRP data preprocessing, the averages computed without the deconvolution approach, and the baseline correction conducted with the deconvolution approach were run in Matlab using Fieldtrip toolbox.

5.3. Data PreProcessing Decisions and Parameters

The aim of this section is to help the reader understand the scripts we have provided and, if they wish, to adapt them for their own purposes. Thus, while we describe the preprocessing decisions taken for our own sample datasets, and the parameters used, alternative decisions may be more appropriate in respect of other data sets. We also note that although in the following sub-sections we only describe the most general decisions taken (in light of space limits), the scripts are supported by extensive comments for each line or part of the code.

5.3.1 EM Data PreProcessing

The EM data that are used in the R script were previously cleaned with SR Research DataViewer such that fixations shorter than 50 ms and longer than 800 ms were excluded from the datasets, and therefore, from the fixation, interest area, saccade, and message reports. In addition, throughout the script (i.e., for each EM report used) we ensure that only experimental trials are included in the dataset for preprocessing.

The cleaned EM datasets are then checked for consecutive fixations, such that observations where there were non-consecutive fixations from the pretarget word to the target word are marked down. Similarly, we mark down any observation where a blink and skip on the pre-target or target words is detected during first-pass reading. Next, we check when the display change occurred and mark down the observations associated with early or late display changes. We define as early the changes that occurred while a fixation was detected on the pre-target word. We classify as late the changes that occurred more than 10 ms ([1second/140 Hz refresh rate] + 3ms) after a fixation was detected on the target word. We use a separate procedure to detect hooks, as they occur during a saccade rather than during a fixation. In this case, we check whether a display update occurred during a saccade that ended on the pre-target word. Finally, we remove from the data all those observations that were marked down as having issues.

In addition, at the end of the R script, we generate a variable to be used later in the analysis to identify the onsets of the interesting fixations in the FRP data, and therefore to distinguish and select the interesting fixation onsets from all the fixation onsets recorded in the EEG data. To create this variable, we compute the time between the first fixation onset on the word of interest in the sentence (i.e., the pretarget and target words) and the time of the trial onset as recorded in the EM data (i.e., TimeLockFixOnset = FixOnset - TTLon). We then add this value to the trial onset as originally recorded in the EEG data (i.e., 'urPresentation'). Indeed, the timings are not zeroed at the beginning of a trial onset in the EEG data. Thus, computing this time is critical to identify the interesting fixation onsets in the EEG data, which will need to be retained for further analyses.

5.3.2. ET-Guided EEG Data PreProcessing

We use the raw EEG data and the event files to run the Matlab code with functions from EEGLAB, EYE-EEG, Unfold and Fieldtrip.

Trial onsets and offsets were recorded as a string of text and represented the number of the trial being displayed in our EM data (e.g., 'TTLon112' and 'TTLoff112' for the onset and offset of trial index 112), while they were recorded as integers and represented the experimental condition of the trial being displayed in the EEG data (e.g., 3 for the onset and 103 for the offset of a trial in the experimental condition 3). Thus, the first step of our Matlab script is to recode these events in order to have the same values for all TTLon/trial onsets (i.e., 601) and all the TTLoff/trial offsets (i.e., 602) of both the EM and EEG recordings in the EEG event structure, regardless of trial index and experimental condition. At this point, we also remove practice and filler trials, and convert the EM raw data (.asc) into Matlab format (.mat) and the EEG raw data (.cnt) into Matlab EEGLAB format (.set files containing meta-information of the data, and .fdt files containing the raw data. Note that .fdt files will be automatically created when saving .set files).

Once the EM and EEG data contain the same value for all trial onsets and the same value for all trial offsets, the data can be synchronised. We use the EYE-EEG extension of EEGLAB for synchronisation, as it allows for importation of the ET data as additional ET data channels in the EEG, synchronisation of EM and EEG data based on common events present in both recordings (i.e., 601 and 602), and estimation of the quality of synchronization between the two recordings.

Following this, we check whether there is any EEG channel with a high percentage of intervals containing 'bad' data. We first use the CRAP algorithm to identify intervals of data where peak-to-peak signal amplitude differences exceed 250 μV within a default time window of 2000 ms. We then calculate the percentage of 'bad' data for each channel, standard score the percentages and examine whether any channel has standard score (z) that is greater than 3. If a channel has z score higher than 3, then that channel is considered 'bad' and data associated with that bad channel are removed from the dataset.

Next, we use the EYE-EEG extension of EEGLAB to apply the filters to the EEG and EOG channels only, excluding the ET channels. We adopt the new basic FIR filter with pass-band edge of 0.1 Hz, and the default transition band width of 0.1 Hz and cutoff frequency of 0.05 Hz (-6dB) as high-pass filter. Our data were already low-pass filtered at 100 Hz, so we do not apply any low-pass filter here. However, it is possible for other researchers to apply a low-pass filter at this point in the pipeline if they have recorded with higher online low-pass filter parameters. The important point here is to include high frequencies that might help optimisation of the ICA decomposition.

Once the EEG data have been synchronised with the ET

data, bad channels detected and the data filtered, we then commence the ICA procedure. First, we detect bad EEG intervals. We use the EYE-EEG extension to identify those EEG intervals of continuous data where the ET channels show extreme values, outside the possible range of the display screen resolution. We then use the CRAP algorithm to identify intervals of data where peak-to-peak signal amplitude differences exceed 250 μV within a default time window of 2000 ms. Here we also check the synchronisation lag between the ET and EEG systems with the cross-correlation procedure included in the EYE-EEG extension. However, the EYE-EEG implementation of the cross-correlation function assumes monopolar EOG recordings, whereas our example dataset used bipolar EOG electrodes. Thus, in the script we use a pair of EEG electrodes from locations that are sensitive for noise from horizontal EMs to approximate the signal content of monopolar horizontal EOGs.

Second, we create optimised ICA training data. Following the OPTICAT procedure (Dimigen, 2020), we filter the data with a high-pass filter of 3 Hz (i.e., to ensure improved ICA modelling performance of corneoretinal dipole and saccadic spike potential) and include frequencies up to approximately 100 Hz (to better model the saccadic spike potential). We then cut the training data into 2s epochs (from trial onset). Next, we cut and mean-center 30 ms epochs (-20 to +10 ms from saccade onset) and append them to the training data. We use 30% of the training data length as length of the appended epochs extracted around saccade onset. Although 100% peri-saccadic sample overweighting allows for increased data on which to run the computations, and thus, better identifies ocular artifacts, we opted for 30% to reduce computational time and resources (see Figure S6, Dimigen, 2020). We next run the extended Infomax ICA algorithm on training data downsampled to 500 Hz to speed up the ICA computations, and then use a threshold of 1.1 in the EYE-EEG extension of EEGLAB to detect the ocular ICs on training data and we plot the results to check them. Finally, we transfer the ICA weights and the identified ocular ICs from the training data to the original filtered (0.1-100 Hz) data.

Based on the previous ICA computation and ocular IC detection, we remove the ocular ICs with the EYE-EEG extension and then low-pass filter the original data at 30 Hz. At this point, if any of the channels was removed from the dataset, we use spherical interpolation to estimate the signal of each 'bad' channel from the neighbouring channels. Following, we use the CRAP algorithm but with stricter values compared to the initial detection of bad EEG intervals within the ICA procedure. That is, we identify intervals of data where amplitude differences exceed 150 μV. Finally, we apply average re-referencing (i.e., all scalp electrodes, no EOG and ET channels) before estimating the FRP data.

5.3.3. FRP Data PreProcessing

At the end of the EM preprocessing procedure, a file was created which included all the necessary information to later conduct statistical analyses on the EM data, as well as a variable called '*TimeLockFixOnset*' with the onset time of all the interesting fixations. At this point in the preprocessing of the EEG data, we import this file in the EEG event structure, to select which fixation is to be used as the time-locking event in the EEG data. This is accomplished by matching the EM ''*TimeLockFixOnset*' values to the unique EEG'*starttime*' values of the fixations. Therefore, by the end of this procedure, we identify the same number of interesting fixations in both the EM and EEG data.

If FRPs are estimated using the deconvolution approach,

we use the Unfold toolbox to build a model. In the Matlab script, we have built a model that includes both type of preview and target frequency as categorical predictors to investigate both their main effects and interaction. In addition, we define the overlap window to be -700 ms to 800 ms around fixations, which we believe provides windows long enough to estimate the activity prior to fixation as well as activity after the fixation. Once the model is solved, we export the betas in the Fieldtrip structure format to be used in Fieldtrip for statistical analysis with cluster-based permutation tests. Finally, we use Fieldtrip functions to baseline the betas with the period of time preceding fixation onset (i.e., -700 to -500 ms for the target word, and -200 to 0 ms for the pretarget word).

If FRPs are estimated without deconvolution, we cut the FRP data into epochs of 1000 ms around each fixation onset (-200 ms to +800 ms), we remove the epochs containing bad data, we baseline the epochs with the period of time (i.e., 200 ms) preceding the fixation onset on the pre-target word for both pre-target and target words, and then compute the average of each condition with Fieldtrip functions to statistically analyse the FRP data with cluster-based permutation tests in Fieldtrip.

Regardless of the FRP estimation approach, we then extract the EM data.

6. Conclusions

In the present paper we have discussed a series of issues and decisions that researchers wishing to start co-registration reading research will need to engage with. Discussion followed the steps of a pipeline we shared, which might be one of the several possible pipelines that may be used when conducting co-registration reading experiments, and in principle, other areas of human cognitive psychology experimentation. Our pipeline was not meant to be prescriptive. On the contrary, each step was designed to be maximally adaptable to the particular experimental situations and data sets that different researchers may encounter or anticipate as they engage in (or consider engaging in) co-registration reading experimentation. The final aim in this paper was to provide some example computational scripts that researchers might find a useful basis for processing co-registration data should they be considering engaging in this kind of research. To provide further guidance, we provided example datasets from a sentence reading experiment, presented the decisions made and parameters used with these datasets, and we provided two scripts (one in R and one in Matlab) that researchers may choose to adapt for their own needs. More generally, our aim in this regard is to encourage and expand the community of researchers using this methodology, which we hope will further current theoretical understanding of the neural and cognitive processes underlying reading, and more general human psychological function.

Contributions

Contributed to conception and design: FD, SPL
Contributed to analysis pipeline design: FD, OL
Contributed to pipeline scripts design: FD, OL
Drafted and/or revised the article: FD, OL, SPL
Approved the submitted version for publication: FD, OL, SPL

Funding Information

Competing Interests

The authors declare no competing interests.

Data Accessibility Statement

Sample raw data and pipeline scripts can be found on this paper's project page on https://osf.io/62bqx/?view_only=d28689a07d7a43bc9f8060bdf2a506a6

REFERENCES

Acunzo, D. J., MacKenzie, G., & van Rossum, M. C. W. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. *Journal of Neuroscience Methods, 209*(1), 212–218. https://doi.org/10.1016/j.jneumeth.2012.06.011

Alday, P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology, 56*(12), e13451. https://doi.org/10.1111/psyp.13451

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial event-related potentials. *Neuropsychologia, 49*(5), 970–983. https://doi.org/10.1016/j.neuropsychologia.2011.01.003

Angele, B., Slattery, T. J., & Rayner, K. (2016). Two stages of parafoveal processing during reading: Evidence from a display change detection task. *Psychonomic Bulletin & Review, 23*(4), 1241–1249. https://doi.org/10.3758/s13423-015-0995-0

Artoni, F., Delorme, A., & Makeig, S. (2018). Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition. *NeuroImage, 175*, 176–187. https://doi.org/10.1016/j.neuroimage.2018.03.016

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baccino, T., & Manunta, Y. (2005). Eye-Fixation-Related Potentials: Insight into parafoveal processing. *Journal of Psychophysiology, 19*(3), 204–215. https://doi.org/10.1027/0269-8803.19.3.204

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37*(1), 13–20. https://doi.org/10.1111/1469-8986.3710013

Bahill, A. T., Clark, M. R., & Stark, L. (1975a). Computer simulation of overshoot in saccadic eye movements. *Computer Programs in Biomedicine, 4*(4), 230–236.

Bahill, A. T., Clark, M. R., & Stark, L. (1975b). Dynamic overshoot in saccadic eye movements is caused by neurological control signal reversals. *Experimental Neurology, 48*(1), 107–122. https://doi.org/10.1016/0014-4886(75)90226-5

Bahill, A. T., & Stark, L. (1975). Neurological control of horizontal and vertical components of oblique saccadic eye movements. *Mathematical Biosciences, 27*(3–4), 287–298. https://doi.org/10.1016/0025-5564(75)90107-8

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Berg, P., & Scherg, M. (1991). Dipole modelling of eye activity and its application to the removal of eye artefacts from the EEG and MEG. *Clinical Physics and Physiological Measurement, 12*(A), 49–54. https://doi.org/10.1088/0143-0815/12/a/010

Bertrand, O., Perrin, F., & Pernier, J. (1985). A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 62*(6), 462–464. https://doi.org/10.1016/0168-5597(85)90058-9

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics, 9*(16), 1–20. https://doi.org/10.3389/fninf.2015.00016

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2017). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology, 55*(e13049), 1–16. https://doi.org/10.1111/psyp.13049

Cardoso, J. F., & Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEE Proceedings F Radar and Signal Processing, 140*(6), 362–370. https://doi.org/10.1049/ip-f-2.1993.0054

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Coco, M. I., Nuthmann, A., & Dimigen, O. (2020). Fixation-related Brain Potentials during Semantic Integration of Object-Scene Information. *Journal of Cognitive Neuroscience, 32*(4), 571–589. https://doi.org/10.1162/jocn_a_01504

Collewijn, H., van der Steen, J., & Steinman, R. M. (1985). Human eye movements associated with blinks and prolonged eyelid closure. *Journal of Neurophysiology, 54*(1), 11–27. https://doi.org/10.1152/jn.1985.54.1.11

Cornelissen, T., Sassenhagen, J., & Võ, M. L.-H. (2019). Improving free-viewing fixation-related EEG potentials with continuous-time regression. *Journal of Neuroscience Methods, 313*, 77–94. https://doi.org/10.1016/j.jneumeth.2018.12.010

Croft, R. J., & Barry, R. J. (2000a). EOG correction: Which regression should we use? *Psychophysiology, 37*(1), 123–125. https://doi.org/10.1111/1469-8986.3710123

Croft, R. J., & Barry, R. J. (2000b). Removal of ocular artifact from the EEG: A review. *Neurophysiologie Clinique/Clinical Neurophysiology, 30*(1), 5–19. https://doi.org/10.1016/s0987-7053(00)00055-1

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience, 10*, 604. https://doi.org/10.3389/fnhum.2016.00604

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research, 1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

de Cheveigné, A., & Simon, J. Z. (2008). Sensor noise suppression. *Journal of Neuroscience Methods, 168*(1), 195–202. https://doi.org/10.1016/j.jneumeth.2007.09.012

Debener, S., Thorne, J., Schneider, T. R., & Viola, F. C. (2010). Using ICA for the analysis of multi-channel EEG data. In M. Ullsperger & S. Debener (Eds.), *Simultaneous EEG and fMRI: Recording, Analysis, and Application: Recording, Analysis, and Application* (pp. 121–134). NY.

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019a). Parafoveal previews and lexical frequency in natural reading: Evidence from eye movements and fixation-related potentials. *Journal of Experimental Psychology: General, 148*(3), 453–474. https://doi.org/10.1037/xge0000494

Degno, F., Loberg, O., Zang, C., Zhang, M., Donnelly, N., & Liversedge, S. P. (2019b). A co-registration investigation of inter-word spacing and parafoveal preview: Eye movements and fixation-related potentials. *PLoS ONE, 14*(12), e0225819. https://doi.org/10.1371/journal.pone.0225819

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Desmedt, J. E., Chalklin, V., & Tomberg, C. (1990). Emulation of somatosensory evoked potential (SEP) components with the 3-shell head model and the problem of 'ghost potential fields' when using an average reference in brain mapping. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 77*(4), 243–258. https://doi.org/10.1016/0168-5597(90)90063-j

Dien, J. (1998). Issues in the application of the average reference: Review, critiques, and recommendations. *Behavior Research Methods, Instruments, & Computers, 30*(1), 34–43. https://doi.org/10.3758/bf03209414

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage, 207*, 116117. https://doi.org/10.1016/j.neuroimage.2019.116117

Dimigen, O., & Ehinger, B. V. (2020). Analyzing combined eye-tracking/EEG experiments with (non) linear deconvolution models. *BioRxiv, 735530.*

Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials. *NeuroImage, 62*(1), 381–393. https://doi.org/10.1016/j.neuroimage.2012.04.006

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General, 140*(4), 552–572. https://doi.org/10.1037/a0023885

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance, 31*(5), 954–969. https://doi.org/10.1037/0096-1523.31.5.954

Ehinger, B. V. (2019). *Unmixed: Linear Mixed Models combined with Overlap Correction for M/EEG analyses. An Extension to the unfold Toolbox.* Conference on Cognitive Computational Neuroscience, Berlin, Germany. https://ccneuro.org/2019/proceedings/0000867.pdf

Ehinger, B. V., & Dimigen, O. (2019). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ, 7*, e7838. https://doi.org/10.7717/peerj.7838

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

Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences*, *103*(18), 7192–7197. https://doi.org/10.1073/pnas.0509557103

Engbert, Ralf, & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/s0042-6989(03)00084-1

Frey, A., Ionescu, G., Lemaire, B., López-Orozco, F., Baccino, T., & Guérin-Dugué, A. (2013). Decision-making in information seeking on texts: An eye-fixation-related potentials investigation. *Frontiers in Systems Neuroscience*, *7*, 39. https://doi.org/10.3389/fnsys.2013.00039

Fröber, K., Stürmer, B., Frömer, R., & Dreisbach, G. (2017). The role of affective evaluation in conflict adaptation: An LRP study. *Brain and Cognition, 116*, 9–16. https://doi.org/10.1016/j.bandc.2017.05.003

Frömer, R., Maier, M., & Abdel Rahman, R. (2018). Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, *12*(48), 1–15. https://doi.org/10.3389/fnins.2018.00048

Frömer, R., Stürmer, B., & Sommer, W. (2016a). The better, the bigger: The effect of graded positive performance feedback on the reward positivity. *Biological Psychology*, *114*, 61–68. https://doi.org/10.1016/j.biopsycho.2015.12.011

Frömer, R., Stürmer, B., & Sommer, W. (2016b). (Don't) Mind the effort: Effects of contextual interference on ERP indicators of motor preparation. *Psychophysiology*, *53*(10), 1577–1586. https://doi.org/10.1111/psyp.12703

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

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

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

Henderson, J. M., Luke, S. G., Schmidt, J., & Richards, J. E. (2013). Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in Systems Neuroscience*, *7*(28), 1–13. https://doi.org/10.3389/fnsys.2013.00028

Hox, J. (2010). *Multilevel Analysis: Techniques and applications* (2nd ed.). Routledge. https://doi.org/10.4324/9780203852279

Hutzler, F., Braun, M., Võ, M. L.-H., Engl, V., Hofmann, M., Dambacher, M., Leder, H., & Jacobs, A. M. (2007). Welcome to the real world: Validating fixation-related brain potentials for ecologically valid settings. *Brain Research*, *1172*(3), 124–129. https://doi.org/10.1016/j.brainres.2007.07.025

Hutzler, F., Fuchs, I., Gagl, B., Schuster, S., Richlan, F., Braun, M., & Hawelka, S. (2013). Parafoveal X-masks interfere with foveal word recognition: Evidence from fixation-related brain potentials. *Frontiers in Systems Neuroscience*, *7*(33), 1–10. https://doi.org/10.3389/fnsys.2013.00033

Hyvärinen, A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. https://www.jstor.org/stable/1390807

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C., & Artieda, J. (2003). Independent component analysis as a tool to eliminate artifacts in EEG: A quantitative study. *Journal of Clinical Neurophysiology*, *20*(4), 249–257. https://doi.org/10.1097/00004691-200307000-00004

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Joyce, C. A., Gorodnitsky, I. F., & Kutas, M. (2004). Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, *41*(2), 313–325. https://doi.org/10.1111/j.1469-8986.2003.00141.x

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

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, *111*(10), 1745–1758. https://doi.org/10.1016/s1388-2457(00)00386-2

Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*(4), 523–532. https://doi.org/10.1111/1469-8986.3740523

Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural Computation*, *9*(7), 1493–1516. https://doi.org/10.1162/neco.1997.9.7.1493

Kapoula, Z. A., Robinson, D. A., & Hain, T. C. (1986). Motion of the eye immediately after a saccade. *Experimental Brain Research*, *61*(2), 386–394. https://doi.org/10.1007/bf00239527

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

Kayser, J., & Tenke, C. E. (2010). In search of the Rosetta Stone for scalp EEG: Converging on reference-free techniques. *Clinical Neurophysiology*, *121*(12), 1973–1975. https://doi.org/10.1016/j.clinph.2010.04.030

Keren, A. S., Yuval-Greenberg, S., & Deouell, L. Y. (2010). Saccadic spike potentials in gamma-band EEG: Characterization, detection and suppression. *Neuroimage*, *49*(3), 2248–2263. https://doi.org/10.1016/j.neuroimage.2009.10.057

Kierkels, J. J. M., Riani, J., Bergmans, J. W. M., & Van Boxtel, G. J. M. (2007). Using an eye tracker for accurate eye movement artifact correction. *IEEE Transactions on Biomedical Engineering*, *54*(7), 1256–1267. https://doi.org/10.1109/tbme.2006.889179

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 238. https://doi.org/10.3389/fpsyg.2010.00238

Kornrumpf, B., Niefind, F., Sommer, W., & Dimigen, O. (2016). Neural correlates of word recognition: A systematic comparison of natural reading and rapid serial visual presentation. *Journal of Cognitive Neuroscience, 28*(9), 1374–1391. https://doi.org/10.1162/jocn_a_00977

Krantz, J. H. (2000). Tell me, what did you see? The stimulus on computers. *Behavior Research Methods, Instruments, & Computers*, *32*(2), 221–229. https://doi.org/10.3758/bf03207787

Kretzschmar, F., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). Parafoveal versus foveal N400s dissociate spreading activation from contextual fit. *NeuroReport*, *20*(18), 1613–1618. https://doi.org/10.1097/wnr.0b013e328332c4f4

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1648–1662. https://doi.org/10.1037/xlm0000128

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

Kristensen, E., Guerin-Dugué, A., & Rivet, B. (2017). Regularization and a general linear model for event-related potential estimation. *Behavior Research Methods*, *49*(6), 2255–2274. https://doi.org/10.3758/s13428-017-0856-z

Kulke, L., & Kulke, V. (2020). Combining eye tracking with EEG: Effects of filter settings on EEG for trials containing task relevant eye-movements. *BioRxiv*. https://doi.org/10.1101/2020.04.22.054882

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. https://doi.org/10.1016/0301-0511(80)90046-0

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*(2), 417–441. https://doi.org/10.1162/089976699300016719

Li, N., Niefind, F., Wang, S., Sommer, W., & Dimigen, O. (2015). Parafoveal processing in reading Chinese sentences: Evidence from event-related brain potentials. *Psychophysiology*, *52*(10), 1361–1374. https://doi.org/10.1111/psyp.12502

Lins, O. G., Picton, T. W., Berg, P., & Scherg, M. (1993). Ocular artifacts in recording EEGs and event-related potentials II: Source dipoles and source components. *Brain Topography*, *6*(1), 65–78. https://doi.org/10.1007/bf01234128

Loberg, O., Hautala, J., Hämäläinen, J. A., & Leppänen, P. H. T. (2018). Semantic anomaly detection in school-aged children during natural sentence reading - A study of fixation-related brain potentials. *PloS One*, *13*(12), e0209741. https://doi.org/10.1371/journal.pone.0209741

Loberg, O., Hautala, J., Hämäläinen, J. A., & Leppänen, P. H. T. (2019). Influence of reading skill and word length on fixation-related brain activity in school-aged children during natural reading. *Vision Research*, *165*, 109–122. https://doi.org/10.1016/j.visres.2019.07.008

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213

López-Peréz, P. J., Dampuré, J., Hernández-Cabrera, J. A., & Barber, H. A. (2016). Semantic parafoveal-on-foveal effects and preview benefits in reading: Evidence from fixation related potentials. *Brain and Language*, *162*, 29–34. https://doi.org/10.1016/j.bandl.2016.07.009

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Maess, B., Schröger, E., & Widmann, A. (2016). High-pass filters and baseline correction in M/EEG analysis. Commentary on: "How inappropriate high-pass filters can produce artefacts and incorrect conclusions in ERP studies of language and cognition." *Journal of Neuroscience Methods*, *266*(15), 164–165. https://doi.org/10.1016/j.jneumeth.2015.12.003

Mahajan, Y., Peter, V., & Sharma, M. (2017). Effect of EEG referencing methods on auditory mismatch negativity. *Frontiers in Neuroscience*, *11*, 560. https://doi.org/10.3389/fnins.2017.00560

Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 145–151). MIT Press.

Makeig, S., & Onton, J. (2009). ERP features and EEG dynamics: An ICA perspective. In S. Luck & E. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 51–86). Oxford.

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

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917. https://doi.org/10.1037/h0037368

Metzner, P., von der Malsburg, T., Vasishth, S., & Rösler, F. (2015). Brain responses to world knowledge violations: a comparison of stimulus-and fixation-triggered event-related potentials and neural oscillations. *Journal of Cognitive Neuroscience*, *27*(5), 1017–1028. https://doi.org/10.1162/jocn_a_00731

Metzner, P., Von Der Malsburg, T., Vasishth, S., & Rösler, F. (2017). The importance of reading naturally: Evidence from combined recordings of eye movements and electric brain potentials. *Cognitive Science*, *41*(S6), 1232–1263. https://doi.org/10.1111/cogs.12384

Meyberg, S., Sommer, W., & Dimigen, O. (2017). How microsaccades relate to lateralized ERP components of spatial attention: A co-registration study. *Neuropsychologia*, *99*, 64–80. https://doi.org/10.1016/j.neuropsychologia.2017.02.023

Mirault, J., Yeaton, J., Broqua, F., Dufau, S., Holcomb, P. J., & Grainger, J. (2020). Parafoveal‐on‐foveal repetition effects in sentence reading: A co‐registered eye‐tracking and electroencephalogram study. *Psychophysiology*, *57*(8), e13553. https://doi.org/10.1111/psyp.13553

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: A review and recommendations. *Frontiers in Human Neuroscience*, *7*(138), 1–11. https://doi.org/10.3389/fnhum.2013.00138

Nakano, T., Kato, M., Morito, Y., Itoi, S., & Kitazawa, S. (2013). Blink-related momentary activation of the default mode network while viewing videos. *Proceedings of the National Academy of Sciences*, *110*(2), 702–706. https://doi.org/10.1073/pnas.1214804110

Niefind, F., & Dimigen, O. (2016). Dissociating parafoveal preview benefit and parafovea-on-fovea effects during reading: A combined eye tracking and EEG study. *Psychophysiology*, *53*(12), 1784–1798. https://doi.org/10.1111/psyp.12765

Nikolaev, A. R., Meghanathan, R. N., & van Leeuwen, C. (2016). Combining EEG and eye movement recording in free viewing: Pitfalls and possibilities. *Brain and Cognition*, *107*, 55–83. https://doi.org/10.1016/j.bandc.2016.06.004

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204. https://doi.org/10.3758/brm.42.1.188

Onton, J., & Makeig, S. (2006). Information-based modeling of event-related brain dynamics. In *Progress in Brain Research* (Vol. 159, pp. 99–120). Elsevier. https://doi.org/10.1016/s0079-6123(06)59007-7

Onton, J., Westerfield, M., Townsend, J., & Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neuroscience & Biobehavioral Reviews*, *30*(6), 808–822. https://doi.org/10.1016/j.neubiorev.2006.06.007

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *156869*, 1–9. https://doi.org/10.1155/2011/156869

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Perrin, F., Pernier, J., Bertnard, O., Giard, M. H., & Echallier, J. F. (1987). Mapping of scalp potentials by surface spline interpolation. *Electroencephalography and Clinical Neurophysiology*, *66*(1), 75–81. https://doi.org/10.1016/0013-4694(87)90141-6

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & 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

Pinherio, J. C., & Bates, D. M. (2004). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag.

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, *6*(278), 1–23. https://doi.org/10.3389/fnhum.2012.00278

Pollatsek, A., Rayner, K., & Balota, D. A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics*, *40*(2), 123–130. https://doi.org/10.3758/bf03208192

Rämä, P. I. A., & Baccino, T. (2010). Eye fixation-related potentials (EFRPs) during object identification. *Visual Neuroscience*, *27*(5–6), 187–192. https://doi.org/10.1017/s0952523810000283

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, *7*(1), 65–81. https://doi.org/10.1016/0010-0285(75)90005-5

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition*, *5*(4), 443–448. https://doi.org/10.3758/bf03197383

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*(1), 21–30. https://doi.org/10.1068/p080021

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The Effects of Frequency and Predictability on Eye Fixations in Reading: Implications for the E-Z Reader Model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720–732. https://doi.org/10.1037/0096-1523.30.4.720

Rayner, K., White, S. J., Kambe, G., Miller, B., & Liversedge, S. P. (2003). On the processing of meaning from parafoveal vision during eye fixations in reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 213–234). Elsevier. https://doi.org/10.1016/b978-044451020-4/50013-x

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157. https://doi.org/10.1037/0033-295x.105.1.125

Richlan, F., Gagl, B., Schuster, S., Hawelka, S., Humenberger, J., & Hutzler, F. (2013). A new high-speed visual stimulation method for gaze-contingent eye movement and brain activity studies. *Frontiers in Systems Neuroscience*, *7*, 24. https://doi.org/10.3389/fnsys.2013.00024

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(4), 515–530. https://doi.org/10.1111/j.1467-9868.2005.00513.x

Rousselet, G. A. (2012). Does filtering preclude us from studying ERP time-courses? *Frontiers in Psychology*, *3*(131), 1–9. https://doi.org/10.3389/fpsyg.2012.00131

Sassenhagen, J., & Draschkow, D. (2019). Cluster‐based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

Schotter, E., & Rayner, K. (2015). The work of the eyes during reading. In A. Pollatsek & R. Treiman (Eds.), *The Oxford Handbook of Reading* (pp. 44–59). Oxford.

Sereno, S. C., & Rayner, K. (2000). The when and where of reading in the brain. *Brain and Cognition*, *42*(1), 78–81. https://doi.org/10.1006/brcg.1999.1167

Sereno, S. C., & Rayner, K. (2003). Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences*, *7*(11), 489–493. https://doi.org/10.1016/j.tics.2003.09.010

Simola, J., Holmqvist, K., & Lindgren, M. (2009). Right visual field advantage in parafoveal processing: Evidence from eye-fixation-related potentials. *Brain and Language*, *111*(2), 101–113. https://doi.org/10.1016/j.bandl.2009.08.004

Slattery, T. J., Angele, B., & Rayner, K. (2011). Eye movements and display change detection during reading. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1924–1938. https://doi.org/10.1037/a0024322

Smeets, J. B. J., & Hooge, I. T. C. (2003). Nature of variability in saccades. *Journal of Neurophysiology*, *90*(1), 12–20. https://doi.org/10.1152/jn.01075.2002

Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*(2), 157–168. https://doi.org/10.1111/psyp.12317

Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169–181. https://doi.org/10.1111/psyp.12320

Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 137–142. https://doi.org/10.3758/bf03204486

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

Urbach, T. P., & Kutas, M. (2006). Interpreting event-related brain potential (ERP) distributions: implications of baseline potentials and variability with application to amplitude normalization by vector scaling. *Biological Psychology*, *72*(3), 333–343. https://doi.org/10.1016/j.biopsycho.2005.12.002

Van der Lubbe, R. H. J., Wauschkuhn, B., Wascher, E., Niehoff, T., Kömpf, D., & Verleger, R. (2000). Lateralized EEG components with direction information for the preparation of saccades versus finger movements. *Experimental Brain Research*, *132*(2), 163–178. https://doi.org/10.1007/s002219900328

VanRullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Frontiers in Psychology*, *2*(365), 1–6. https://doi.org/10.3389/fpsyg.2011.00365

Volkmann, F. C., Riggs, L. A., & Moore, R. K. (1980). Eyeblinks and visual suppression. *Science*, *207*(4433), 900–902. https://doi.org/10.1126/science.7355270

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Wang, C.-A., & Inhoff, A. W. (2013). Extraction of linguistic information from successive words during reading: Evidence for spatially distributed lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 662–677. https://doi.org/10.1037/a0030242

Weiss, B., Knakker, B., & Vidnyánszky, Z. (2016). Visual processing during natural reading. *Scientific Reports*, *6*(26902), 1–16. https://doi.org/10.1038/srep26902

Widmann, A., & Schröger, E. (2012). Filter effects and filter artifacts in the analysis of electrophysiological data. *Frontiers in Psychology*, *3*(233), 1–5. https://doi.org/10.3389/fpsyg.2012.00233

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data - a practical approach. *Journal of Neuroscience Methods*, *250*, 34–46. https://doi.org/10.1016/j.jneumeth.2014.08.002