Fixation-related brain potentials during semantic integration of object-scene information

Moreno I. Coco [1,2], Antje Nuthmann[3] and Olaf Dimigen[4]

1. School of Psychology, The University of East London, London, UK
2. Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal
3. Institute of Psychology, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
4. Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

This research was supported by the Leverhulme Trust (grant ECF-014-205) and Fundação para a Ciência e Tecnologia (grant PTDC/PSI-ESP/30958/2017) to M. I. C, while he was a Research Fellow at the University of Edinburgh. The authors thank Benedikt Ehinger for helpful discussions on EEG deconvolution techniques. Please address correspondence to moreno.cocoi@gmail.com or olaf.dimigen@hu-berlin.de.

Abstract

In vision science, a particularly controversial topic is whether and how quickly the semantic information about objects is available outside foveal vision. Here, we aimed at contributing to this debate by co-registering eye-movements and EEG while participants viewed photographs of indoor scenes that contained a semantically consistent or inconsistent target object. Linear deconvolution modelling was used to analyse the event-related potentials (ERP) evoked by scene onset as well as the fixation-related potentials (FRPs) elicited by the fixation on the target object (*t*) and by the preceding fixation (*t*-1). Object-scene consistency did not influence the probability of immediate target fixation or the ERP evoked by scene onset, which suggests that object-scene semantics was not accessed immediately. However, during the subsequent scene exploration, inconsistent objects were prioritized over consistent objects in extrafoveal vision (i.e., looked at earlier) and were more effortful to process in foveal vision (i.e., looked at longer). In FRPs, we demonstrate a fixation-related N300/N400 effect, whereby inconsistent objects elicit a larger frontocentral negativity than consistent objects. In line with the behavioural findings, this effect was already seen in FRPs aligned to the pre-target fixation *t*-1 and persisted throughout fixation *t*, indicating that the extraction of object semantics can already begin in extrafoveal vision. Taken together, the results emphasize the usefulness of combined EEG/eye-movement recordings for understanding the mechanisms of object-scene integration during natural viewing.

Keywords: object-scene integration; foveal and peripheral vision; semantic processing; fixation-related potentials, eye tracking, N300/N400, regression-ERPs

Introduction

In our daily activities, for example when we search for something in a room, our attention is mostly oriented to objects. The time course of object recognition and the role of overt attention in this process are therefore topics of considerable interest in the visual sciences. In the context of real-world scene perception, the question of what constitutes an object is a more complex question than intuition would suggest (e.g., Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011). An object is likely a hierarchical construct (e.g., Feldman, 2003), with both low-level features (e.g., visual saliency) and high-level properties (e.g., semantics) contributing to its identity. Accordingly, when a natural scene is inspected with eye-movements, the observer's attentional selection is thought to be based either on objects (e.g., Nuthmann & Henderson, 2010), image features (saliency; Itti, Koch, & Niebur, 1998) or some combination of the two (e.g., Stoll, Thrun, Nuthmann, & Einhäuser, 2015).

An early and uncontroversial finding is that the recognition of objects is mediated by their semantic consistency. For example, an object that the observer would not expect to occur in a particular scene (e.g., a toothbrush in a kitchen) is recognized less accurately (e.g., Biederman, 1972; Davenport & Potter, 2004; Fenske, Aminoff, Gronau, & Bar, 2006) and looked at for longer that an expected object (e.g., Cornelissen & Võ, 2017; De Graef, Christiaens, & D'Ydewalle, 1990; Henderson, Weeks Jr, & Hollingworth, 1999).

What is more controversial, however, is the exact time course along which the meaning of an object is processed and how this semantic processing then influences the overt allocation of visual attention (see Wu, Wick, & Pomplun, 2014, for a review). Two interrelated questions are at the core of this debate: (1) How much time is needed to access the meaning of objects after a scene is displayed and (2) can object semantics be extracted before the object is overtly attended, that is, while the object is still outside of high-acuity foveal vision (> 1° eccentricity) or even in the periphery (> 5° eccentricity).

Evidence that the meaning of not-yet-fixated objects can capture overt attention comes from experiments that have used sparse displays of several standalone objects (e.g., Belke, Humphreys, Watson, Derrick, Meyer, & Telling, 2008; Cimminella, Della Sala & Coco, in press; Moores, Laiti, & Chelazzi, 2003; Nuthmann, de Groot, Huettig, & Olivers, 2019). For example, across three different experiments Nuthmann et al. found that the very first saccade in the display was directed more frequently to objects that were semantically related to a target object rather than to unrelated objects.

Whether such findings generalize to objects embedded in real-world scenes is currently an open research question. The size of the visual span – that is the area of the visual field from which observers can take in useful information (see Rayner, 2014 for review) – is large in scene viewing. For object-in-scene search, it corresponded to approximately 8º in each direction from fixation (Nuthmann, 2013). This opens up the possibility that both low-level and high-level object properties can be processed outside the fovea. This is the case for low-level visual features: objects that are highly salient (i.e., visually distinct) are preferentially selected for fixation (e.g., Stoll et al., 2015). But what about high-level semantic information? If extra-foveal semantic processing takes place, then objects that are inconsistent with the scene context (which are also thought to be more informative, Antes, 1974) should be fixated earlier in time than consistent ones (Loftus & Mackworth, 1978; Mackworth & Morandi, 1967).

However, results from eye-movement studies on this issue have been mixed. A number of studies have indeed reported evidence for an inconsistent object advantage (e.g., Bonitz & Gordon, 2008; Borges, Fernandes, & Coco, 2019; LaPointe & Milliken, 2016; Loftus & Mackworth, 1978; Underwood, Templeman, Lamming, & Foulsham, 2008). Among these studies, only Loftus & Mackworth (1978) have reported evidence for *immediate* extrafoveal attentional capture (i.e. within the first fixation) by object-scene semantics. In this

study, which used relatively sparse line drawings of scenes, the mean amplitude of the saccade into the critical object was more than 7°, suggesting that viewers could process semantic information based on peripheral information obtained in a single fixation. Several other studies, however, have failed to find any advantage for inconsistent objects in attracting overt attention (e.g., De Graef, Christiaens, & D'Ydewalle, 1990; Henderson, Weeks, & Hollingworth, 1999; Võ & Henderson, 2009, 2011). In these experiments, only measures of foveal processing – such as gaze duration – were influenced by object-scene consistency, with longer fixations times on inconsistent than on consistent objects.

Interestingly, a similar controversy exists in the literature on eye guidance in sentence reading. Although some degree of parafoveal processing during reading is uncontroversial, it is less clear whether semantic information is acquired from the parafovea (Andrews & Veldre, 2019, for review). Most evidence from studies involving readers of English has been negative (e.g., Rayner, Balota, & Pollatsek, 1986), whereas results from reading German (e.g., Hohenstein & Kliegl, 2014) and Chinese (e.g., Yan, Richter, Shu, & Kliegl, 2009) suggest that parafoveal processing can advance up to the level of semantic processing.

The processing of object-scene inconsistencies and its time course have also been investigated in electrophysiological studies (e.g., Ganis & Kutas, 2003; Mudrik, Lamy, & Deouell, 2010). In event-related potentials (ERPs), it is commonly found that scene-inconsistent objects elicit a larger negative brain response compared to consistent ones. This long-lasting negative shift typically starts as early as 200-250 ms after stimulus onset (e.g., Mudrik, Shalgi, Lamy, & Deouell, 2014; Draschkow, Heikel, Võ, Fiebach, & Sassenhagen, 2018) and has its maximum at frontocentral scalp sites, in contrast to the centroparietal N400 effect for words (e.g., Kutas & Federmeier, 2011). The effect was found for objects that appeared at a cued location after the scene background was already shown (Ganis & Kutas, 2003), for objects that were photoshopped into the scene (Mudrik, Lamy, & Deouell, 2010;

Mudrik, et al., 2014; Coco, Araujo, & Petersson, 2017), and for objects that were part of realistic photographs (Võ & Wolfe, 2013). These ERP effects of object-scene consistency have typically been subdivided into two distinct components: N300 and N400. The earlier part of the negative response, usually referred to as N300, has been taken to reflect the context-dependent difficulty of object identification, whereas the later N400 has been linked to semantic integration processes after the object is identified (e.g., Dyck & Brodeur, 2015). The present study was not designed to differentiate between these two subcomponents, especially considering that their scalp distribution is strongly overlapping or even topographically indistinguishable (Draschkow et al., 2018). Thus, for reasons of simplicity, we will in most cases simply refer to all frontocentral negativities as "N400".

One limiting factor of existing ERP studies is that the data were gathered using steady-fixation paradigms in which the free exploration of the scene through eye-movements was not permitted. Instead, the critical object was typically large and/or located relatively close to the centre of the screen, and ERPs were time-locked to the onset of the image (e.g., Mudrik et al., 2010). Due to these limitations, it remains unclear whether foveation of the object is a necessary condition for processing object-scene consistencies, or whether such processing can at least begin in extrafoveal vision.

In the current study, we used fixation-related potentials (FRPs), that is EEG waveforms aligned to fixation onset, to shed new light on the controversial findings of the role of foveal versus extrafoveal vision in extracting object semantics, while providing insights into the patterns of brain activity that underlie them (for reviews about FRPs see Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011; Nikolaev, Meghanathan, & van Leeuwen, 2016).

FRPs have previously been used to investigate the brain-electric correlates of natural reading, as opposed to serial word presentation, helping researchers to provide finer details

about the online processing of linguistic features (such as word predictability, Kretzschmar, Bornkessel-Schlesewsky, & Schlesewsky, 2009; Kliegl, Dambacher, Dimigen, Jacobs, & Sommer, 2012) or the dynamics of the perceptual span during reading (e.g., parafovea-on-fovea effects, Niefind & Dimigen, 2016). More recently, the co-registration method has also been applied to investigate active visual search (e.g., Devillez, Guyader, & Guerin-Dugue, 2015; Kamienkowski, Ison, Quiroga, & Sigman, 2012; Kaunitz et al., 2014), object identification (Rämä & Baccino, 2010), and affective processing in natural scene viewing (Simola, Le Fevre, Torniainen, & Baccino, 2015).

In the present study, we simultaneously recorded eye-movements and FRPs during the viewing of real-world scenes to distinguish between three alternative hypotheses on object-scene integration that can be derived from the literature: (A) one glance of the scene is sufficient to extract object semantics from extrafoveal vision (e.g., Loftus & Mackworth, 1978), (B) extrafoveal processing of object-scene semantics is possible but takes some time to unfold (e.g., Bonitz & Gordon, 2008; Underwood et al., 2008), and (C) the processing of object semantics requires foveal vision, that is, a direct fixation of the critical object (e.g., De Graef et al., 1990; Henderson et al., 1999; Võ & Henderson, 2009). We note that these possibilities are not mutually exclusive, an issue we elaborate on in the *General Discussion*.

For the behavioural data, these hypotheses translate as follows: under (A), the probability of immediate target fixation should reveal that already the first saccade on the scene goes more often towards inconsistent than consistent objects. Under (B), there should be no effect on the first eye-movement, but the latency to first fixation on the critical object should be shorter for inconsistent than consistent objects. Under (C), only fixation times on the critical object itself should differ as a function of object-scene consistency, with longer gaze durations on inconsistent objects.

For the electrophysiological data analysis, we used a novel regression-based analysis approach (linear deconvolution modelling; Cornelissen, Sassenhagen, & Võ, 2019, Dandekar, Privitera, Carney, & Klein, 2011; Dimigen & Ehinger, 2019; Ehinger & Dimigen, 2018; Smith & Kutas, 2015b), which allowed us to control for the confounding influences of overlapping potentials and oculomotor covariates during natural viewing, which can otherwise distort the neural responses. In the EEG, hypothesis (A) can be tested by computing the ERP time-locked to the onset of the scene on the display, following the traditional approach. Given that the critical objects in our study were not placed directly in the centre of the screen from which observers started their exploration of the scene, any effect of object-scene congruency in this ERP would suggest that object semantics is rapidly processed in extrafoveal vision, even before the first eye-movement is generated, in line with Loftus & Mackworth, 1978. Under hypothesis (B) we would not expect to see an effect in the scene-onset ERP. Instead, we should find a negative brain potential (N400) for inconsistent as compared to consistent objects in the FRP aligned to the fixation that *precedes* the one that first lands on the critical object. Finally, if (C) is correct, an N400 for inconsistent objects should only arise once the critical object is foveated, i.e., in the FRP aligned to the target fixation (fixation *t*). In contrast, no consistency effects should appear in the scene-onset ERP or in the FRP aligned to the pre-target fixation (fixation *t*-1). To preview the results, both the eye-movement as well as the EEG data lend support for hypothesis (B).

Methods

Design and task overview

We designed a short-term visual working memory change detection task, illustrated in Figure 1 and 2. During the *study phase*, participants were exposed to photographs of indoor scenes (e.g., a bathroom), each of which contained a target object that was either semantically

192 *consistent* (e.g., toothpaste) or *inconsistent* (e.g., a flashlight) with the scene context. In the
193 following *recognition phase*, after a short retention interval of 900 ms, the same scene was
194 shown again, but in half of the trials either the identity, the location, or both the identity and
195 location of the target object had changed relative to the study phase.

196 Insert Figure 1 and 2 about here

197 The participants' task was to indicate with a keyboard press whether or not a change had
198 happened to the scene (see also LaPointe & Milliken, 2016). All eye-movement and EEG
199 analyses in the present article focus on the semantic consistency manipulation of the target
200 object during the *study phase*.

201 ***Participants***

202 Twenty-four participants (9 male) between the ages of 18 and 33 (mean: 25.0 years)
203 took part in the experiment after providing written informed consent. They were compensated
204 with £7 per hour. All participants had normal or corrected-to-normal vision. Data from an
205 additional two participants was recorded but removed from the analysis due to excessive
206 scalp muscle (EMG) activity or skin potentials in the raw EEG. Ethics approval was obtained
207 from the Psychology Research Ethics Committee of the University of Edinburgh.

208 ***Apparatus and Recording***

209 Scenes were presented on a 19" CRT monitor (Iiyama Vision Master Pro 454) at a
210 vertical refresh rate of 75 Hz. At the viewing distance of 60 cm, each scene subtended 35.8º ×
211 26.9º (width × height). Eye-movements were recorded monocularly from the dominant eye
212 using an SR Research EyeLink 1000 desktop-mounted system at a sampling rate of 1000 Hz.
213 Eye dominance for each participant was determined with a parallax test. A chin and forehead
214 rest was used to stabilize the participant's head. Nine-point calibrations were run at the

beginning of each session and whenever the participant's fixation deviated by > 0.5° horizontally or > 1° vertically from a drift correction point presented at trial onset.

The EEG was recorded from 64 active electrodes at a sampling rate of 512 Hz using BioSemi ActiveTwo amplifiers. Four electrodes, located near the left and right canthus and above and below the right eye, recorded the electro-oculogram (EOG). All channels were referenced against the BioSemi common mode sense (CMS; active electrode) and grounded to a passive electrode. The BioSemi hardware is DC coupled and applies digital low-pass filtering through the A/D-converter's decimation filter, which has a 5th order sinc response with a -3 dB point at 1/5th of the sample rate (corresponding approximately to a 100 Hz low-pass filter).

Offline, the EEG was re-referenced to the average of all scalp electrodes and filtered using EEGLAB's (Delorme & Makeig, 2004) Hamming-windowed sinc FIR filter (pop_eegfiltnew.m) with default settings. The lower edge of the filter's passband was set to 0.2 Hz (with -6 dB attenuation at 0.1 Hz) and the upper edge to 30 Hz (with -6 dB attenuation at 33.75 Hz). Eye tracking and EEG were synchronized using shared triggers sent via the parallel port of the stimulus presentation PC to the two recording computers. Synchronization was performed offline using the EYE-EEG extension (v0.8) for EEGLAB (Dimigen et al., 2011). All datasets were aligned with a mean synchronization error ≤ 2 ms as computed based on trigger alignment after synchronization.

Materials & Rating

Stimuli consisted of 192 colour photographs of indoor scenes (e.g., bedrooms, bathrooms, offices). Real target objects were placed in the physical scene, before each picture was taken with a tripod under controlled lighting conditions and with a fixed aperture (i.e., there was no photo-editing). One scene is shown in Figure 1; miniature version of all stimuli used in the present study are found as part of the *Supplementary Materials*. Of the 192 scenes,

96 were conceived as change items and 96 as no-change items. Each one of the 96 change scenes was created in four versions. In particular, each scene (e.g., a bathroom) was photographed with two alternative target objects in it, one that was consistent with the scene context (e.g., a toothbrush) and one that was not (e.g., a flashlight). Moreover, each of these two objects was placed at two alternative locations (left or right side) within the scene (e.g., either on the sink or on the bathtub). Accordingly, three types of change were implemented during the recognition phase (*Congruency*, *Location,* and *Both*); see *Procedure* section below.

Each of the 96 no-change scenes was also a real photograph with either a consistent or an inconsistent object in it, which was again located in either the left or right half of the scene. Across the 96 no-change scenes, factors consistency (consistent vs. inconsistent objects) and location (left and right) were also balanced. However, each no-change scene was unique, that is, we did not create four different versions of each no-change scene. The data of the 96 no-change scenes, which were originally conceived to be filler trials, was included to improve the signal-to-noise ratio of the EEG analyses, as these scenes also had a balanced distribution of inconsistent and consistent objects.

As explained above, each scene contained a critical object that was either consistent or inconsistent with the scene context. Object consistency was assessed in a pre-test rating study by eight naïve participants who were not involved in any other aspect of the study. Each participant rated all of the no-changes scenes as well as one of the four versions of each change-scene (counterbalanced across raters). Together with each scene, raters saw a box with a cropped image of the critical object. They were asked (a) to write down the name for the displayed object, and (b) to respond to the question "*How likely is it that this object would be found in this room?*" using a six-point Likert scale (1-6). For the object naming, a mean naming agreement of 96.35% was obtained. Furthermore, consistent objects were judged as significantly more likely (mean = 5.78, SD = ± 0.57) to appear in the scene than inconsistent

objects (1.88 ± 1.11), as confirmed by an independent-samples Kruskal-Wallis *H*-test ($\chi^2(1) = 616.09$, $p < .001$).

In addition, we ensured that there was no difference between consistent and inconsistent objects on three important low-level variables: object size (pixels square), distance from the centre of the scene (degrees of visual angle) and mean visual saliency of the object as computed using the Adaptive Whitening Saliency model (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012). Table 1 provides additional information about the target object. Paired *t*-tests showed no significant difference between consistency conditions in object size, $t(476) = -1.2$, $p = 0.2$; visual saliency, $t(476) = 0.1$, $p = 0.9$; and distance from the centre, $t(476) = 0.48$, $p = 0.6$.

The position of each target object was marked with an invisible rectangular bounding box, which was used to implement the gaze contingency mechanism (described in the *Procedure* section below) and to determine whether a fixation was inside the target object. The average width of the bounding box was 6.1° ± 2.0 for consistent objects and 6.1° ± 2.1 for inconsistent objects (see Table 1); the average height was 5.1° ± 1.8 and/or 5.4° ± 2.2, respectively. The average distance of the object centroid from the centre of the scene was 12.1° (± 2.8) for consistent and 11.7° (± 3.0) for inconsistent objects.

Procedure

A schematic representation of the task is shown in Figure 2. Each trial started with a drift correction of the eye-tracker. Afterwards, the study scene was presented (e.g., a bathroom). The display duration of the study scene was controlled by a gaze-contingent mechanism that ensured that participants fixated the target object (e.g., toothbrush or flashlight) at least once during the trial. Specifically, the study scene disappeared on average 2000 ms (with a random jitter of ±200 ms, drawn from a uniform distribution) after the participant's eyes left the invisible bounding box of the target object (and provided that the

target had been fixated for at least 150 ms). The jittered delay of about 2000 ms was implemented to prevent participants from learning to associate the last fixated object during the study phase with the changed object during the recognition phase. If the participant did not fixate the target object within 10 s, the study scene disappeared from the screen and the retention interval was triggered, which lasted for 900 ms.

In the following recognition phase (data not analysed here), the scene was presented again, either with (50% of trials) or without (50% of trials) a change to an object in the scene. Three types of object changes occurred with equal probability: Location, Consistency, or Both. In the (a) *Location* condition**,** the target object changed its position and moved either from left to right or from right to left to another plausible location within the scene (e.g., a toothbrush was placed elsewhere within the scene). In the (b) *Consistency* condition, the object remained in the same location, but was replaced with another object of opposite semantic consistency (e.g., the toothbrush was replaced by a flashlight or vice-versa). Finally, in the (c) *Both* condition, the object was both replaced and moved within the scene (e.g., a toothbrush was replaced by a flashlight at a different location).

During the recognition phase, participants had to indicate whether they noticed any kind of change within the scene by pressing the arrow keys on the keyboard. Afterwards, the scene disappeared and the next trial began. If participants did not respond within 10 s, a missing response was recorded.

The type of change between trials was fully counterbalanced using a Latin Square rotation. Specifically, the 96 change trials were distributed across 12 different lists, implementing the different types of change. This implies that each participant was exposed to an equal number of consistent and inconsistent change trials. The 96 no-change trials also comprised an equal number of consistent and inconsistent scenes and were the same for each participant. All 192 trials were presented in a randomized order. These trials were preceded

by four practice trials at the start of the session. Written instructions were given to explain the task which took 20-40 minutes to complete. The experiment was implemented using the SR Research Experiment Builder software.

Data preprocessing

Eye-movement events and data exclusion

Fixations and saccades events were extracted from the raw gaze data using the SR Research Data Viewer software, which performs saccade detection based on velocity and acceleration thresholds of 30° s^{-1} and 9,500° s^{-2}, respectively. To provide directly comparable results for eye-movement behaviour and FRP analyses, we discarded all trials on which we did not have clean data from both recordings. Specifically, from a total of 4,608 trials (24 participants × 192 trials), we excluded 10 trials (0.2%) because of machine error (i.e., no data was recorded for those trials), 689 trials (15.0%) because the participant responded incorrectly after the recognition phase and 494 trials (10.7%) because the target object was not fixated during the study phase. Finally, we removed an additional 97 trials (2.1%) for which the target fixation overlapped with intervals of the EEG that contained non-ocular artefacts (see below). The final dataset therefore comprised 3,318 unique trials: 1,567 for the consistent condition and 1,751 for the inconsistent condition. Per participant, this corresponded to an average of 65.3 trials (± 6.9, range = 48-78) for consistent and 73.0 trials (± 6.9, range = 59-82) for inconsistent items. Due to the fixation check, participants were always fixating at the screen centre when the scene appears on the display. This on-going central fixation was removed from all analyses.

EEG ocular artefact correction

EEG recordings during free viewing are contaminated by three types of ocular artefacts (e.g., Dimigen, 2018; Plöchl, Ossandón, & König, 2012) which need to be removed to get at the genuine brain activity. Here we applied an optimized variant (Dimigen, 2018) of

Independent Component Analysis (ICA, Jung et al., 1998), which uses the information provided by the eye-tracker to objectively identify ocular ICA components (Plöchl et al., 2012).

In a first step, we created optimized ICA training data by high-pass filtering a copy of the EEG at 2 Hz (Winkler, Debener, Müller, & Tangermann, 2015; Dimigen, 2018) and segmenting it into epochs lasting from scene onset until 3 s thereafter. This high pass-filtered training data was entered into an extended Infomax ICA using EEGLAB, and the resulting unmixing weights were then transferred to the original (i.e. less strictly filtered) recording (Viola, Debener, Thorne, & Schneider, 2010). From this original EEG dataset, we then removed all independent components whose time course varied more strongly during saccade intervals (defined as lasting from -20 ms before saccade onset until 20 ms after saccade offset) than during fixations with the threshold for the variance ratio (saccade/fixation, see Plöchl et al., 2012) set to 1.3. The artefact-corrected continuous EEG was then back-projected to the sensor space. For a validation of the ICA procedure, please refer to Supplementary Figure S1.

In a next step, intervals with residual non-ocular artefacts (e.g., EMG bursts) were detected by shifting a 2000 ms moving window in steps of 100 ms across the continuous recording. Whenever the voltages within the window exceeded a peak-to-peak threshold of 100 µV in at least one of the channels, all data within the window was marked as “bad” and subsequently excluded from analysis. Within the linear deconvolution framework (see below), this can easily be done by setting all predictors to zero during these bad EEG intervals (Smith & Kutas, 2015b), meaning that the data in these intervals will not affect the computation.

Analysis

Eye-movement data

Dependent measures: Behavioural analyses focused on four eye-movement measures commonly reported in the semantic consistency literature: (a) cumulative probability of having fixated the target object as a function of the ordinal fixation number, (b) the probability of immediate object fixation, (c) the latency to first fixation on the target object, and (d) the gaze duration on the target object (cf. Võ & Henderson, 2009).

Linear-mixed effect modelling: Eye-movement data were analysed using linear mixed-effects models (LMM) and generalized linear mixed-effects models (GLMM) as implemented in the `lme4` package in `R` (Bates et al., 2015). The only exception was the cumulative probability of first-fixations on the target for which a generalized linear model (GLM) was used. One advantage of (G)LMM modelling is that it allows one to simultaneously model the intrinsic variability of both participants and scenes (e.g., Nuthmann & Einhäuser, 2015).

In all analyses, the main predictor was the *Consistency* of the critical object (contrast coding: Consistent = -0.5, Inconsistent = 0.5) in the study scene. In the (G)LMMs, *Participant* (24) and *Scene* (192) were included as random intercepts[1]. The cumulative probability of object fixation was analysed using a GLM with a binomial (probit) link. This model included the *Ordinal Number of Fixation* on the scene as a predictor; it was entered as a continuous variable ranging from 1 to a maximum of 28 (the 99th quantile).

In the tables of results, we report the beta coefficients, *t*-values (LMM), *z*-values (GLMM), and *p*-values for each model. The level of significance was calculated from an *F*-test

[1] We did not include random slopes for two reasons: For *Participant*, the inclusion of a random slope led to a small variance and perfect correlation between intercept and slope. For the random effect *Scene*, only the change trials were fully counterbalanced in terms of location and consistency, meaning that the slope for *Consistency* could not be estimated for the no-change trials.

based on the Satterthwaite approximation to the effective degrees of freedom (Satterthwaite, 1946), where *p*-values in GLMMs are based on asymptotic Wald tests.

Electrophysiological data

Linear Deconvolution Modelling (first level of analysis): EEG measurements during active vision are associated with two major methodological problems: overlapping potentials and low-level signal variability (Dimigen & Ehinger, 2019). Overlapping potentials arise from the rapid pace of active information sampling through eye-movements, which causes the neural responses that are evoked by subsequent fixations on the stimulus to overlap with each other. Because the average fixation duration usually varies between conditions, this changing overlap can easily confound the measured waveforms. A related issue is the mutual overlap between the ERP elicited by the initial presentation of the stimulus and the FRPs evoked by the subsequent fixations on it. This second type of overlap is especially important in experiments like ours, in which the critical fixations occurred at different latencies after scene onset in the two experimental conditions.

The problem of signal variability refers to the fact that low-level visual and oculomotor variables can also influence the morphology of the predominantly visually-evoked fixation-related neural responses (e.g., Dimigen et al., 2011; Kristensen, Rivet, & Guerin-Dugué, 2017; Nikolaev et al., 2016). The most relevant of these variables, which is known to modulate the entire FRP waveform, is the amplitude of the saccade that precedes fixation onset (e.g., Dandekar et al., 2011; Thickbroom, Knezevič, Carroll, & Mastaglia, 1991). One option for controlling the effect of saccade amplitude is to include it as a continuous covariate in a massive univariate regression-model (Smith & Kutas, 2015a, 2015b), in which a separate regression model is computed for each EEG time point and channel (Weiss, Knakker & Vidnyánszky, 2016). However, this method does not account for overlapping potentials.

409 An approach that allows one to simultaneously control for overlapping potentials and
410 low-level covariates is deconvolution within the linear model (for tutorial reviews see
411 Dimigen & Ehinger, 2019; Smith & Kutas, 2015a, 2015b), sometimes also called continuous-
412 time regression (Smith & Kutas, 2015b). Initially developed to separate overlapping BOLD
413 responses (e.g., Serences, 2004), linear deconvolution has also been applied to separate
414 overlapping potentials in ERP (Smith & Kutas, 2015b) and FRP paradigms (Cornelissen et
415 al., 2019; Dandekar et al., 2011; Ehinger & Dimigen, 2018; Kristensen, et al. 2017). Another
416 elegant property of this approach is that the ERPs elicited by scene onset and the FRPs
417 elicited by fixations on the scene can be disentangled and simultaneously estimated in the
418 same regression model. The benefits of deconvolution are illustrated in more detail in
419 Supplementary Figures S2 and S3.
420 Here, we applied this technique by using the new *unfold* toolbox (Ehinger & Dimigen,
421 2018), which represents the first-level analysis and provides us with the partial effects (i.e.
422 the beta coefficients or "regression-ERPs", Smith & Kutas, 2015a, 2015b) for each predictor
423 of interest. In a first step, both stimulus onset events and fixation onset events were included
424 as stick functions (also called finite impulse responses, FIR) in the design matrix of the
425 regression model. To account for overlapping activity from adjacent experimental events, the
426 design matrix was then time-expanded in a time window between -300 and +800 ms around
427 each stimulus and fixation onset event. Time-expansion means that the time points within this
428 window are added as predictors to the regression model. Because the temporal distance
429 between subsequent events in the experiment is variable, it is possible to disentangle their
430 overlapping responses. Time-expansion with stick functions is explained in Serences (2004)
431 and Ehinger & Dimigen (2018, see their Figure 2). The model was run on EEG data sampled
432 at the original 512 Hz, that is, no down-sampling was performed.
433 Using Wilkinson notation, the model formula for scene onset events was defined as:

```
ERP ~ 1 + Consistency
```

In this formula, the beta coefficients for the intercept (`1`) capture the shape of the overall waveform of the stimulus-ERP in the consistent condition, which was used as the reference level, whereas those for *Consistency* capture the differential effect of presenting an inconsistent object in the scene (relative to a consistent object) on the ERP. The coefficients for the predictor *Consistency* are therefore analogous to a difference waveform in a traditional ERP analysis (Smith & Kutas, 2015a, 2015b) and would reveal if semantic processing already occurs immediately after the initial presentation of the scene.

In the same regression model, we also included the onsets of all fixations made on the scene. Fixation onsets were modelled with the formula

```
FRP ~ 1 + Consistency * Type + Sacc_Amplitude
```

Thus, we predicted the FRP for each time-point as a function of the semantic *Consistency* of the target object (Consistent vs. Inconsistent; Consistent as reference level) in interaction with the *Type* of fixation (Critical fixation vs. Non-target fixation; Non-target fixation as reference level). In this model, any FRP consistency effects elicited by the pre-target or target fixation would appear as an interaction between *Consistency* and fixation *Type*. In addition, we included the incoming *Saccade Amplitude* (in degrees of visual angle) as a continuous linear covariate to control for the effect of saccade size on the FRP waveform[2]. Thus, the full model was as follows:

[2] Other low-level variables, such as local image features in the currently foveated image region (e.g., luminance, spatial frequency) are also known to modulate the FRP waveform. In the model presented here, we did not include these other covariates because (1) their influence on the FRP waveform is small compared to that of saccade amplitude and (2) the properties of the target object (such as its visual saliency) did not differ between the two levels of object consistency (see *Materials and Rating*). For reasons of simplicity, saccade amplitude was included as a linear predictor in the current model, although its influence on the FRP becomes non-linear for large saccades (e.g., Dandekar et al., 2011)..

```
{ERP ~ 1 + Consistency,
FRP ~ 1 + Consistency * Type + Sacc_Amplitude}
```

This regression model was then solved for the betas using MATLAB's *glmfit* solver (without regularization).

The deconvolution model specified by the formula above was run twice: in one version, we treated the pre-target fixation (*t*-1) as the critical fixation, in the other version the target fixation (*t*). In a given model, all fixations but the critical ones were defined as non-target fixations. FRPs for fixation *t*-1 and for fixation *t* were estimated in two separate runs of the model, rather than simultaneously within the same model, because the estimation of overlapping activity was much more stable in this case. In other words, while the deconvolution method allowed us to control for much of the overlapping brain activity from other fixations, we were not able to use the model to directly separate the (two) N400 consistency effects elicited by the fixations *t*-1 and *t* [3].

However, virtually identical results were obtained when we included it as a non-linear (spline) predictor instead (Dimigen & Ehinger, 2019).

[3] In theory, a more elegant model would include *Type* as a three-level predictor, with levels *pre-target*, *target*, and *non-target fixation*. In principle, this would allow us to dissociate which parts of the N400 consistency effects are elicited by fixation *t*-1 versus fixation *t*. The practical disadvantage of this approach is that the overlapping activities from both *t*-1 and *t* would then be estimated on comparatively fewer observations (compared to the extremely stable estimate for the numerous non-target fixations). This is critical because compared to the limited amount of jitter in natural fixation durations, N400 effects are a long-lasting response, which makes the deconvolution more challenging. Specifically, we found that with the three-level model, model outputs became extremely noisy and did not yield significant consistency effects for any EEG time-locking point. By defining either fixation *t*-1 or fixation *t* as the critical fixation in two separate runs of the model, and by treating all other fixations as non-target fixations, the estimation becomes very robust. This simpler model still removes most of the overlapping activity from other fixations. However, the consistency-specific activity evoked by fixation *t*-1 (i.e., the N400 effect) will not be removed from the FRP aligned to the fixation *t* and vice versa.

Both runs of the model (the one for *t*-1 and *t*) also yield an estimate for the scene onset-ERP, but because the results for the scene-onset ERP were virtually identical, we present the betas from the first run of the model.

The average number of events entering the model per participant was 65.7 and 73.6 for scene onsets (consistent and inconsistent condition, respectively), 864.2 and 887.9 for non-target fixations (*nt*), 58.3 and 59.8 for pre-target fixations (*t*-1), and 63.8 and 70.6 for target fixations (*t*).

Baseline placement for FRPs: Another challenging issue for free-viewing EEG experiments is the choice of an appropriate neutral baseline interval for the FRP waveforms (Dimigen et al., 2011; Nikolaev et al., 2016). Baseline placement is particularly relevant for experiments on extrafoveal processing where we do not know in advance when EEG differences will arise, and whether they may already develop prior to fixation onset.

For the pre-target fixation *t*-1 and non-target fixations *nt*, we used a standard baseline interval by subtracting the mean channel voltages between -200 and 0 ms before the event (note that the saccadic spike potential ramping up at the end of this interval was almost completely removed by our ICA procedure; see Figure A1). For fixation *t*, we cannot use such a baseline because semantic processing may already be ongoing by the time the target object is fixated. Thus, to apply a neutral baseline to fixation *t*, we subtracted the mean channel voltages in the 200 ms interval before the *preceding* fixation *t*-1 also from the FRP aligned to the target fixations *t* (see Nikolaev et al., 2016 for similar procedures). The scene-onset ERP was corrected with a standard pre-stimulus baseline (-200 to 0 ms).

Group statistics for EEG (second level of analysis): To perform second-level group statistics, averaged EEG waveforms at the single-subject level ("regression-ERPs") were reconstructed from the beta coefficients of the linear deconvolution model. These regression-based ERPs are directly analogous to subject-level averages in a traditional ERP analysis

(Smith & Kutas, 2015a). We then used two complementary statistical approaches to examine consistency effect in the EEG: linear mixed models and a cluster-based permutation test.

LMM in a-priori defined time windows. LMM were used to provide hypothesis-based testing motivated by existing literature. Specifically, we adopted the spatiotemporal definitions by Võ & Wolfe (2013) and compared the consistent and inconsistent condition in the time windows from 250 – 350 ms (early effect) and 350 – 600 ms (late effect) at a mid-central region-of-interest (ROI) of nine electrodes (comprising FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, and CP2). Because the outputs provided by the linear deconvolution model (the first-level analysis) are already aggregated at the level of subject-averages, the only predictor included in these LMMs was the *Consistency* of the object. Furthermore, to minimize the risk of Type I error (Barr, Levy, Scheepers, & Tily, 2013) we started with a random effect structure with *Participant* as random intercept and slope for the *Consistency* predictor. This random effect structure was then evaluated and backwards-reduced using the `step` function of the `lmerTest` package (Kuznetsova, Brockhoff, & Christensen, 2017) to retain the model that was justified by the data, i.e., it converged, and it was parsimonious in the number of parameters (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

Cluster permutation tests. It is still largely unknown to what extent the topography of traditional ERP effects translates to natural viewing. Therefore, in order to test for consistency effects across all channels and time points, we additionally applied the Threshold-Free Cluster Enhancement (TFCE) procedure developed by Smith & Nichols (2009) and adapted for EEG data by Mensen & Khatami (2013, http://github.com/Mensen/ept_TFCE-matlab). In a nutshell, TFCE is a non-parametric permutation test that controls for multiple comparisons across time and space, while maintaining relatively high sensitivity (e.g. compared to a Bonferroni correction). Its advantage over previous cluster permutation tests (e.g., Maris & Oostenveld, 2007) is that it

517 does not require the experimenter to set an arbitrary cluster-forming threshold. In the first
518 stage of the TFCE procedure, a raw statistical measure (here: *t*-values) is weighted according
519 to the support provided by clusters of similar values at surrounding electrodes and time
520 points. In the second stage, these cluster-enhanced *t*-values are then compared to the
521 maximum cluster-enhanced values observed under the null hypotheses (based on n=2000
522 random permutations of the data). In the present manuscript (Figures 4 and 5), we not only
523 report the global result of the test, but also plot the spatiotemporal extent of the first-stage
524 clusters, since they provide some indication about which time points and electrodes likely
525 contributed to the overall significant effect established by the test. Please note, however, that
526 unlike the global test result, these first-stage values are not stringently controlled for false
527 positives and do not establish precise effect onset or offsets (Sassenhagen & Draschkow,
528 2019). We report them here as a descriptive statistic.

529 Insert Figure 3 and Tables 1, 2, 3 about here

530 Finally, for purely descriptive purposes and to provide a-priori information for future
531 studies, we also plot the 95% between-subject confidence interval for the consistency effects
532 at the central ROI (corresponding to sample-by-sample paired *t*-testing without correction for
533 multiple comparisons; see also Mudrik et al., 2014) in Figures 4 and 5.

534 **Results**

535 ***Task performance (change detection task)***

536 Following the recognition phase, participants pressed a button to indicate whether or
537 not a change had taken place within the scene. Response accuracy in this task was high (mean
538 = 85.0%± 35.7%) and did not differ as a function of whether the study scene contained a
539 consistent (84.6%± 36.1%) or an inconsistent (85.3%± 35.3%) target object.

Eye-movement behaviour

Figure 3A shows the cumulative probability of having fixated the target object as a function of the ordinal number of fixation and semantic consistency, and Table 2 reports the corresponding GLM model coefficients. We found a significant main effect of *Consistency*; overall, inconsistent objects were looked at with a higher probability than consistent objects. As expected, the cumulative probability of looking at the critical object increased as a function of the *Ordinal Number of Fixation*. There was also a significant interaction between the two variables.

Complementing this global analysis, we analysed the very first eye-movement during scene exploration to assess whether observers had immediate extrafoveal access to object-scene semantics (Loftus & Mackworth, 1978). The mean probability of immediate object fixation was 12.77%; we observed a numeric advantage of inconsistent objects over consistent objects (Figure 3B) but this difference was not significant (Table 3). The latency to first fixation on the target object is another measure to capture the potency of an object in attracting early attention in extrafoveal vision (e.g., Underwood & Foulsham, 2006; Võ & Henderson, 2009). This measure is defined as the time elapsed between the onset of the scene image and the first fixation on the critical object. Importantly, this latency was significantly shorter for inconsistent as compared to consistent objects (Figure 3C, Table 3).

Moreover, we analysed gaze duration as a measure of foveal object processing time (e.g., Henderson et al., 1999). First-pass gaze duration for a critical object is defined as the sum of all fixation durations from first entry to first exit. On average, participants looked longer at inconsistent (520 ms) than consistent objects (409 ms) before leaving the target object for the first time, and this difference was significant (Table 3). Table 1 summarizes additional oculomotor characteristics in the two conditions of object consistency.

565 Supplementary Figures S4 and S5 visualize the locations of the pre-target, target, and
566 post-target fixations for two example scene stimuli.

567 Insert Figure 4, 5, 6 and Table 4 here

568 ***Electrophysiological results***

569 Figures 4 and 5 depict the ERP evoked by the presentation of the scene as well as the
570 FRPs for the three types of fixation that were analysed. Results focus on the mid-central ROI
571 for which effects of object-scene consistency have been previously reported. Waveforms for
572 other scalp sites are depicted in Supplementary Figures S6 to S9.

573 ***Scene-onset ERP.*** The left panels of Figure 4 show the grand-average ERP aligned to
574 scene onset. Although inspection of the scalp maps indicated slightly more positive
575 amplitudes over central right-hemispheric electrodes in the inconsistent condition, these
576 differences were not statistically significant. Specifically, no effect of *Consistency* was found
577 with the LMM analysis in the early or the late time window (see Table 4 for detailed LMM
578 results). Similarly, the TFCE test across all channels and time points yielded no significant
579 *Consistency* effect (all p-values > 0.2, see Figure 4D). Thus, the semantic consistency of the
580 target object did not influence the neural response to the initial presentation of the scene.

581 ***Non-target fixation, nt.*** Next, we tested whether fixations on scenes with an
582 inconsistent object evoke a globally different neural response than those on scenes containing
583 a consistent object. As the right panels of Figure 4 show, this was not the case: Consistency
584 had no effect on the FRP for non-target (*nt*) fixations, neither in the LMM analysis (see Table
585 4) nor in the TFCE statistic (all p-values > 0.2, see Figure 4H).

586 ***Pre-target fixation, t-1.*** Figure 5 depicts the FRPs aligned to the pre-target and target
587 fixations. Importantly, in the FRP aligned to the pre-target fixation t-1, waveforms began to
588 clearly diverge between the two consistency conditions, developing into a long-lasting fronto-
589 central negativity in the inconsistent as compared to the consistent condition (Figure 5A and

590 5B; see also Figure A6). The scalp distribution of this difference, shown in Figure 6, closely
591 resembled the frontocentral N400 (and N300) previously reported in ERP studies on object-
592 scene consistency (e.g., Mudrik et al., 2014; Võ & Wolfe, 2013). In the LMM analyses
593 conducted on the mid-central ROI, this effect was marginally significant ($p < 0.1$) for the
594 early time window (250 to 350 ms), but became highly significant between 350 and 600 ms
595 ($p < 0.001$, Table 4). The TFCE test across all channels and time points revealed a significant
596 effect of consistency on the pre-target FRP ($p < 0.05$). Figure 5C also shows the extents of the
597 underlying spatiotemporal clusters, computed in the first stage of the TFCE procedure.
598 Between 372 ms and 721 ms after fixation onset, we observed a cluster of 14 frontocentral
599 electrodes that was shifted slightly to the left hemisphere. This N400 modulation on the pre-
600 target fixation could be seen even in traditionally-averaged FRP waveforms without any
601 control of overlapping potentials (see Supplementary Figure S3). In summary, we were able
602 to measure a significant frontocentral N400 modulation during natural scene viewing that
603 already emerged in FRPs aligned to the pre-target fixation.

604 On average, the target fixation *t* occurred at a median latency of 240 ms (± 18 ms)
605 after fixation *t*-1, as marked by the vertical dashed line in Figure 5B. If we take the extent of
606 the cluster from the TFCE tests as a rough approximation for the likely onset of the effect in
607 the FRP, this means that, on average, at the time when the ERP consistency effect started
608 (372 ms) the eyes had been looking at the target object for only 132 ms (372 minus 240 ms).

609 ***Target fixation, t.*** An anterior N400 effect was also clearly visible in the FRP aligned
610 to fixation *t.* In the LMM analysis at the central ROI, the effect was significant in both the
611 early (250-350 ms, $p < 0.01$) and late window (350-600 ms, $p < 0.05$; see Table 4). However,
612 compared to the effect aligned to the pre-target fixation, this N400 was significant at only a
613 few electrodes in the TFCE statistic (Cz, FCz, and FC1; see Figure 6). Aligned to the target
614 fixation *t*, the N400 also peaked extremely early, with the maximum of the difference curve

already observed at 200 ms after fixation onset (Figure 5G). Qualitatively, a frontocentral negativity was already visible much earlier than that, within the first 100 ms after fixation onset (Figure 5D). The TFCE permutation test confirmed an overall effect of consistency ($p < 0.05$) on the target-locked FRP. Figure 5G also shows the extents of the underlying first-stage clusters. For the target fixation, clusters only extended across a brief interval between 151 and 263 ms after fixation onset, an interval during which the N400 effect also reached its peak.

Figure 5E shows that, numerically, voltages at the central ROI were more negative in the inconsistent condition during the baseline interval already, that is, before the critical object was fixated. To understand the role of activity already present before fixation onset, we repeated the FRP analyses for fixation *t* after applying a standard baseline correction, with the baseline placed immediately before the target fixation itself (-200 to 0 ms). This way, we eliminate any weak N400-like effects that may have already been on-going before target fixation onset. Interestingly, in the resulting FRP waveforms, the target-locked N400 effects were weakened: The N400 effect now failed to reach significance in the TFCE statistic and in the LMM analysis for the second window (350 to 600 ms; see last row of Table 4) and only remained significant for the early window (250 to 350 ms). This indicates that some N400-like negativity was already ongoing before target fixation onset. To summarize, we found no immediate influences of object-scene consistency in ERPs time-locked to scene onset. However, N400 consistency effects were found in FRPs aligned to the target fixation (*t*) and in those aligned to the pre-target fixation (*t*-1).

Discussion

Substantial research in vision science has been devoted to understanding the behavioural and neural mechanisms underlying object recognition (e.g., Biederman, 1972; Loftus & Mackworth, 1978). At the core of this debate are the type of object features that are accessed (e.g., low-level vs. high-level), the time-course of their processing (e.g., pre-

attentive vs. attentive), and the region of the visual field in which these features can be acquired (e.g., foveal vs. extrafoveal). A particularly controversial topic is whether and how quickly the semantic properties of objects are available outside foveal vision.

In the current study, we approached these questions from a new perspective by co-registering eye-movements and EEG while participants freely inspected images of real-world scenes in which a critical object was either consistent or inconsistent with the scene context. As a novel finding, we demonstrate a fixation-related N400 effect during natural scene viewing. Moreover, behavioural and electrophysiological measures converge to suggest that the extraction of object-scene semantics can already begin in extrafoveal vision, before the critical object is fixated.

It is a rather undisputed finding that inconsistent objects, such as a flashlight in a bathroom, require increased processing when selected as targets of overt attention. Accordingly, several eye-movement studies have reported longer gaze durations on inconsistent than consistent objects, probably reflecting the greater effort required to resolve the conflict between object meaning and scene context (e.g., Cornelissen & Võ, 2017; De Graef et al., 1990; Henderson et al., 1999). In addition, a number of traditional ERP studies using steady-fixation paradigms have found that inconsistent objects elicit a larger negative brain response at frontocentral channels (an N300/N400 complex) as compared to consistent objects (e.g., Coco et al., 2017; Ganis & Kutas, 2003; Mudrik et al., 2010).

However, previous research with eye-movements remained inconclusive on whether semantic processing can take place prior to foveal inspection of the object. Evidence in favour of extrafoveal processing of object-scene semantics comes from studies in which inconsistent objects were selected for fixation earlier than consistent ones (e.g., Borges et al., 2019; LaPointe & Milliken, 2016; Underwood et al., 2008). However, other studies have not found evidence for earlier selection of inconsistent objects (e.g., De Graef et al., 1990;

Henderson et al., 1999; Võ & Henderson, 2009, 2011). Extrafoveal and peripheral vision are known to be crucial for saccadic programming (e.g., Nuthmann, 2014). Therefore, any demonstration that semantic information can act as a source of guidance for fixation selection in scenes implies that some semantic processing must have occurred prior to its fixation, that is, in extrafoveal vision.

ERPs are highly sensitive to semantic processing (Kutas & Federmeier, 2011) and provide excellent temporal resolution to investigate the time course of object processing. However, an obvious limitation of existing ERP studies is that observers were not allowed to explore the scene with saccadic eye-movements, thereby constraining their normal attentional dynamics. Instead, the critical object was usually large and/or placed near the point of fixation. Hence, these studies were unable to establish whether semantic processing can take place prior to its foveal inspection.

In the current study, we addressed this problem by simultaneously recording behavioural and brain-electric correlates of object processing. Specifically, we analysed different eye-movement responses that tap into extrafoveal and foveal processing along with FRPs time-locked to the first fixation on the critical object *(t*) and the fixation preceding it (*t*-1). We also analysed the scene-onset ERP evoked by the trial-initial presentation of the image. Recent advances in linear deconvolution methods for EEG (e.g., Ehinger & Dimigen, 2018) allowed us to disentangle the overlapping brain potentials produced by the scene onset and the subsequent fixations, and to control for the modulating influence of saccade amplitude on the FRP.

The eye-movement behaviour showed no evidence for hypothesis (A) as outlined in the *Introduction*, according to which semantic information can exert an immediate effect on eye-movement control (Loftus & Mackworth, 1978). Specifically, the mean probability of immediate object fixation was fairly low (12.8%) and not modulated by *Consistency*. Instead,

690 the data lends support to hypothesis (B) according to which extrafoveal processing of object-
691 scene semantics is possible but takes some time to unfold. In particular, the results for the
692 latency to first fixation of the critical object show that inconsistent objects were, on average,
693 looked at sooner than consistent objects (cf. Bonitz & Gordon, 2008; Underwood et al.,
694 2008). At the same time, we observed longer gaze durations on inconsistent objects,
695 replicating previous findings (e.g., De Graef et al., 1990; Henderson et al., 1999; Võ &
696 Henderson, 2009). Thus, we found behavioural evidence for the extrafoveal processing of
697 object-scene (in)consistencies, but also differences in the subsequent foveal processing.
698 The question then remains why existing eye-movement studies have provided very
699 different results, ranging from rapid processing of semantic information in peripheral vision
700 to a complete lack of evidence for extrafoveal semantic processing. Researchers have
701 suggested that the outcome may depend on factors related to the critical object or the scene in
702 which it is located. Variables that may (or may not) facilitate the appearance of the
703 incongruency effect include visual saliency (e.g., Henderson et al., 1999; Underwood &
704 Foulsham, 2006), image clutter (Henderson & Ferreira, 2004), and the critical object's size
705 and eccentricity (Gareze & Findlay, 2007). Therefore, an important question for future
706 research is to identify the specific conditions under which extrafoveal semantic information
707 can be extracted, or when the three outlined hypotheses and/or outcomes would prevail.
708 Returning to the present data, the FRP waveforms showed a negative shift over frontal
709 and central scalp sites when participants fixated a scene-inconsistent object. This result is in
710 agreement with traditional ERP studies that have shown an frontocentral N300/N400
711 complex after passive foveal stimulation (e.g., Coco et al., 2017; Ganis & Kutas, 2003;
712 Mudrik et al., 2014; Võ & Wolfe, 2013) and extends this finding for the first time to a natural
713 viewing situation with eye-movements. Regarding the time course, the present data suggest

that the effect was already initiated during the preceding fixation (*t*-1), but then carried on through fixation (*t*) on the target object.

As a cautionary note, we emphasize that it is not trivial to unambiguously ascribe typical N400 (and N300) effects in the EEG to either extrafoveal or foveal processing. The reason is that these canonical congruency effects only begin 200-250 ms after stimulus onset (Draschkow et al., 2018; Mudrik et al., 2010). This means that even a purely extrafoveal effect would be almost impossible to measure during the pre-target fixation (*t*-1) itself, since it would only emerge at a time when the eyes are already moving to the target object. That being said, three properties of the observed FRP consistency effect suggest that it was already initiated during the pre-target fixation:

First, due to the temporal jitter introduced by variable fixation durations, an effect that only arises in foveal vision should be the most robust in the FRP averages aligned to fixation *t*, but latency-jittered and attenuated in those aligned to fixation *t*-1. However, the opposite was the case: At least qualitatively, a frontocentral N400 effect was seen at more electrodes (Figure 6) and for longer time intervals (Figure 5) in the FRP aligned to the pre-target fixation as compared to the actual target fixation. The second argument for extrafoveal contributions to the effect is the forward-shift in its time course. Relative to fixation *t*, the observed N400 occurred almost instantly: As the effect topographies in Figure 5H show, the frontocentral negativity for inconsistent objects was qualitatively visible within the first 100 ms after fixation onset and the effect reached its peak after just 200 ms. Clusters underlying the TFCE test were also restricted to an early time range between 151 and 263 ms after fixation onset and therefore to a much earlier interval to what we would expect from the canonical N300 or N400 effect elicited by foveal stimulation.

Of course, it is possible that even purely foveal N400 effects may emerge earlier during active scene exploration with eye movements as compared to the latencies established

739 in traditional ERP research. For example, it is reasonable to assume that during natural vision,
740 observers pre-process some low-level (non-semantic) features of the soon-to-be fixated object
741 in extrafoveal vision (cf. Nuthmann, 2017). This non-semantic preview benefit might then
742 speed up the timeline of foveal processing (including the latency of semantic access) once the
743 object is fixated (cf. Dimigen, et al., 2012, for reading). Moreover, if eye movements are
744 permitted, observers have more time to build a representation of the scene before they foveate
745 the target, and this increased contextual constraint may also affect the N400 timing (but see
746 Kutas & Hillyard, 1984). Importantly, however, neither of these two accounts could explain
747 why the N400 effect is stronger – rather than much weaker – in the waveforms aligned to
748 fixation *t*-1 as compared to fixation *t*. The fact that the eye movement data also provided
749 clear evidences in favour of extra-foveal processing further strengthens our interpretation of
750 the N400 timing.

751 Finally, we found that the N400 consistency effect aligned to the target fixation (*t*)
752 became weaker (and non-significant in two out of the three statistical measures considered) if
753 the baseline interval for the FRP analysis was placed directly before this target fixation.
754 Again, this indicates that at least a weak frontocentral negativity in the inconsistent condition
755 was already present during the baseline period before the target was fixated. Together, these
756 results are difficult to reconcile with a pure foveal processing account and are more consistent
757 with the notion that semantic processing of the object was at least initiated in extrafoveal
758 vision (and then continued after it was foveated).

759 Crucially, we did not find any effect of target consistency in the traditional ERP
760 aligned to scene onset. In line with the behavioural results, this goes against the most extreme
761 hypothesis A postulating that object semantics can be extracted from peripheral vision
762 already at the first glance of a scene (Loftus & Mackworth, 1978). Similarly, there was no
763 effect of consistency on the FRPs evoked by the non-target fixations on the scene (Figure 4);

this was also the case in a control analysis that only included non-target fixations that occurred earlier than *t*-1 and at an extrafoveal distance between 3º and 7º from the target object (see Supplementary Figure S10). All these analyses suggest that the semantic information of the critical object started during fixation *t*-1. However, from any given fixation there are many candidate locations that could potentially be chosen for the next saccade (cf. Tatler, Brockmole, & Carpenter, 2017). Thus, it is conceivable that observers may have partially acquired semantic information of the critical object outside foveal vision prior to fixation *t*-1, but without selecting it as a saccade target. Such reasoning leaves open the possibility that observers may have already picked up some information about the target object's semantics during these occasions.

Taken together, our behaviour and electrophysiological findings are consistent with the claim formulated in hypothesis B that objects can be recognized outside of the fovea or even in the visual periphery, at least to some degree. Indirectly, our results also speak to the debate about the unit of saccade targeting and, by inference, attentional selection during scene viewing. Finding effects of object-scene semantics on eye guidance is evidence in favour of object- and meaning-based, rather than image-based guidance of attention in scenes (e.g., Hwang, Wang, & Pomplun, 2011; Henderson, Hayes, Peacock, & Rehrig, 2019).

In sum, our findings converge to suggest that the visual system is capable of accessing semantic features of objects in extrafoveal vision to guide attention towards objects that do not fit to the scene's overall meaning. They also highlight the utility of investigating attentional and neural mechanisms in parallel to uncover the mechanisms underlying object recognition during the unconstrained exploration of naturalistic scenes.

788

789 **References**

790 Andrews, S., & Veldre, A. (2019). What is the most plausible account of the role of parafoveal processing in
791 reading? *Language and Linguistics Compass*, *13*(7), e12344.
792 Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62.
793 https://doi.org/10.1037/h0036799
794 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis
795 testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.
796 https://doi.org/10.1016/j.jml.2012.11.001
797 Bates, D., Mächler, M., Bolker, B. M., Walker, S. C., Machler, M., Bolker, B. M., … Walker, S. C. (2015).
798 Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
799 https://doi.org/10.18637/jss.v067.i01
800 Belke, E., Humphreys, G. W., Watson, Derrick, G., Meyer, A. S., & Telling, A. L. (2008). Top-down effects of
801 semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception &*
802 *Psychophysics*, *70*(8), 1444–1458. https://doi.org/10.3758/PP
803 Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77–80.
804 https://doi.org/10.1126/science.177.4043.77
805 Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene
806 viewing. *Acta Psychologica*, *129*(2), 255–263. https://doi.org/10.1016/j.actpsy.2008.08.006
807 Borges, M. T., Fernandes, E. G., & Coco, M. I. (2019). Age-related differences during visual search: the role of
808 contextual expectations and cognitive control mechanisms. *Aging, Neuropsychology, and Cognition*.
809 https://doi.org/10.1080/13825585.2019.1632256
810 Brouwer, A.-M., Reuderink, B., Vincent, J., van Gerven, M. A. J., & van Erp, J. B. F. (2013). Distinguishing
811 between target and nontarget fixations in a visual search task using fixation-related potentials. *Journal of*
812 *Vision*, *13*(3), :17, 1-17. https://doi.org/10.1167/13.3.17
813 Cimminella, F., Della Sala, S., & Coco, M. I. (in press). Extra-foveal Processing of Object Semantics Guides
814 Early Overt Attention During Visual Search. *Attention, Perception, & Psychophysics*.
815 https://doi.org/10.3758/s13414-019-01906-1
816 Coco, M. I., Araujo, S., & Petersson, K. M. (2017). Disentangling stimulus plausibility and contextual
817 congruency: Electro-physiological evidence for differential cognitive dynamics. *Neuropsychologia*, *96*,
818 150–163. https://doi.org/10.1016/j.neuropsychologia.2016.12.008

819 Cornelissen, T. H. W., Sassenhagen, J., & Võ, M. L. H. (2019). Improving free-viewing fixation-related EEG
820 potentials with continuous-time regression. *Journal of Neuroscience Methods*, *313*, 77–94.
821 https://doi.org/10.1016/j.jneumeth.2018.12.010
822 Cornelissen, T. H. W., & Võ, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-scene
823 inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics*, *79*(1), 154–
824 168. https://doi.org/10.3758/s13414-016-1203-7
825 Dandekar, S., Privitera, C., Carney, T., & Klein, S. A. (2011). Neural saccadic response estimation during
826 natural viewing. *Journal of Neurophysiology*, *107*(6), 1776–1790. https://doi.org/10.1152/jn.00237.2011
827 Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological*
828 *Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x
829 De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object
830 identification. *Psychological Research*, *52*(4), 317–329. https://doi.org/10.1007/BF00868064
831 Delorme, A., & Makeig, S. (2004). EEGLAB: an open sorce toolbox for analysis of single-trail EEG dynamics
832 including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.
833 https://doi.org/10.1016/j.jneumeth.2003.10.009
834 Devillez, H., Guyader, N., & Guerin-Dugue, A. (2015). An eye fixation-related potentials analysis of the P300
835 potential for fixations onto a target object when exploring natural scenes. *Journal of Vision*, *15*((13); 20),
836 1–31. https://doi.org/10.1167/15.13.20
837 Dimigen, O. (2018). Optimized ICA-based removal of ocular EEG artifacts from free viewing experiments.
838 *BioRxiv*, (0), 446955. https://doi.org/10.1101/446955
839 Dimigen, O., & Ehinger, B. (2019). Analyzing combined eye-tracking/EEG experiments with (non)linear
840 deconvolution models. *BioRxiv*. https://doi.org/10.1101/735530
841 Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: A
842 study with fixation-related brain potentials. *NeuroImage*, *62*(1), 381–393.
843 https://doi.org/10.1016/j.neuroimage.2012.04.006
844 Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of Eye Movements
845 and EEG in Natural Reading: Analyses and Review. *Journal of Experimental Psychology: General*,
846 *140*(4), 552–572. https://doi.org/10.1037/a0023885
847 Draschkow, D., Heikel, E., Võ, M. L.-H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence from MVPA
848 for different processes underlying the N300 and N400 incongruity effects in object-scene processing.

Neuropsychologia, *120*, 9–17. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2018.09.016

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

Ehinger, B., & Dimigen, O. (2018). Unfold: An integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *BioRxiv*, 360156. https://doi.org/10.1101/360156

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252–256. https://doi.org/10.1016/S1364-6613(03)00111-6

Fenske, M. J., Aminoff, E., Gronau, N., & Bar, M. (2006). Chapter 1 Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in Brain Research*, *155 B*, 3–21. https://doi.org/10.1016/S0079-6123(06)55001-0

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

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, *30*(1), 51–64. https://doi.org/10.1016/j.imavis.2011.11.007

Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. P. G. van Gompel & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford: Elsevier.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York: Psychology Press.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), 19. https://doi.org/10.3390/vision3020019

Henderson, J. M., Weeks, Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037//0096-1523.25.1.210

879 Hohenstein, S., & Kliegl, R. (2014). Semantic preview benefit during reading. *Journal of Experimental*
880 *Psychology: Learning Memory and Cognition*, *40*(1), 166–190. https://doi.org/10.1037/a0033670
881 Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes.
882 *Vision Research*, *51*(10), 1192–1205. https://doi.org/10.1016/j.visres.2011.03.010
883 Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE*
884 *Transaction on Pattern Analysis and Machine Learning*, *20*(11), 1254–1259.
885 https://doi.org/10.1109/TPAMI.2012.125
886 Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (1998).
887 Extended ICA removes artifacts from electroencephalographic recordings. *Advances in Neural*
888 *Information Processing Systems*, *10*, 894–900.
889 Kamienkowski, J. E., Ison, M. J., Quiroga, R. Q., & Sigman, M. (2012). Fixation-related potentials in visual
890 search: A combined EEG and eye tracking study. *Journal of Vision*, *12*(7), :4, 1-20.
891 https://doi.org/10.1167/12.7.4
892 Kaunitz, L. N., Kamienkowski, J. E., Varatharajah, A., Sigman, M., Quiroga, R. Q., & Ison, M. J. (2014).
893 Looking for a face in the crowd: Fixation-related potentials in an eye-movement visual search task.
894 *NeuroImage*, *89*, 297–305. https://doi.org/10.1016/j.neuroimage.2013.12.006
895 Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., & Sommer, W. (2012). Eye movements and brain
896 electric potentials during reading. *Psychological Research*. https://doi.org/10.1007/s00426-011-0376-x
897 Kretzschmar, F., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). Parafoveal versus foveal N400s
898 dissociate spreading activation from contextual fit. *NeuroReport*, *20*(18), 1613–1618.
899 https://doi.org/10.1097/WNR.0b013e328332c4f4
900 Kristensen, E., Guerin-Dugué, A., & Rivet, B. (2017). Regularization and a general linear model for event-
901 related potential estimation. *Behavior Research Methods*, *49*(6), 2255–2274.
902 https://doi.org/10.3758/s13428-017-0856-z
903 Kristensen, E., Rivet, B., & Guerin-Dugué, A. (2017). Estimation of overlapped Eye Fixation Related Potentials:
904 The General Linear Model, a more flexible framework than the ADJAR algorithm. *Journal of Eye*
905 *Movement Research*, *10*(1), 1–27. https://doi.org/http://dx.doi.org/10.16910/jemr.10.1.7
906 Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of
907 the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647.
908 https://doi.org/10.1146/annurev.psych.093008.131123

909 Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed
910 Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13
911 LaPointe, M. R. P., & Milliken, B. (2016). Semantically incongruent objects attract eye gaze when viewing
912 scenes for change. *Visual Cognition*, *24*(1), 63–77. https://doi.org/10.1080/13506285.2016.1185070
913 Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing.
914 *Journal of Experimental Psychology. Human Perception and Performance*, *4*(4), 565–572.
915 https://doi.org/10.1037/0096-1523.4.4.565
916 Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects information details within picutres. *Perception*
917 *and Psychophysics*, *2*(11), 547–552. https://doi.org/10.3758/BF03210264
918 Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of*
919 *Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
920 Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in
921 linear mixed models. *Journal of Memory and Language*, *94*, 305–315.
922 https://doi.org/10.1016/j.jml.2017.01.001
923 Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-
924 parametric statistics. *NeuroImage*, *67*, 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027
925 Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective
926 attention. *Nature Neuroscience*, *6*(2), 182–189. https://doi.org/10.1038/nn996
927 Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous
928 object-scene processing. *Neuropsychologia*, *48*(2), 507–517.
929 https://doi.org/10.1016/j.neuropsychologia.2009.10.011
930 Mudrik, L., Shalgi, S., Lamy, D., & Deouell, L. Y. (2014). Synchronous contextual irregularities affect early
931 scene processing: Replication and extension. *Neuropsychologia*, *56*, 447–458.
932 https://doi.org/10.1016/j.neuropsychologia.2014.02.020
933 Niefind, F., & Dimigen, O. (2016). Dissociating parafoveal preview benefit and parafovea-on-fovea effects
934 during reading: A combined eye tracking and EEG study. *Psychophysiology*, *53*(12), 1784–1798.
935 https://doi.org/10.1111/psyp.12765
936 Nikolaev, A. R., Meghanathan, R. N., & van Leeuwen, C. (2016). Combining EEG and eye movement recording
937 in free viewing: Pitfalls and possibilities. *Brain and Cognition*, *107*, 55–83.
938 https://doi.org/10.1016/j.bandc.2016.06.004

939 Nuthmann, A. (2013). On the visual span during object search in real-world scenes. *Visual Cognition*, *21*(7),
940 803–837. https://doi.org/10.1080/13506285.2013.832449
941 Nuthmann, A. (2014). How do the regions of the visual field contribute to object search in real-world scenes?
942 Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and*
943 *Performance*, *40*(1), 342–360. https://doi.org/10.1037/a0033854
944 Nuthmann, A., de Groot, F., Huettig, F., & Olivers, C. N. L. (2019). Extrafoveal attentional capture by object
945 semantics. *Plos One*, *14*(5), e0217051. https://doi.org/10.1371/journal.pone.0217051
946 Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation
947 selection in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 82–96.
948 https://doi.org/10.1111/nyas.12705
949 Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of*
950 *Vision*, *10*(8), :20, 1-19. https://doi.org/10.1167/10.8.20
951 Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: identification,
952 characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in*
953 *Human Neuroscience*, *6*(October), 1–23. https://doi.org/10.3389/fnhum.2012.00278
954 Rämä, P., & Baccino, T. (2010). Eye fixation-related potentials (EFRPs) during object identification. *Visual*
955 *Neuroscience*, *27*(5–6), 187–192. https://doi.org/10.1017/S0952523810000283
956 Rayner, K. (2014). The gaze-contingent moving window in reading: Development and review. *Visual Cognition*,
957 *22*(3), 242–258. https://doi.org/10.1080/13506285.2013.879084
958 Rayner, K., Balota, D. A., & Pollatsek, A. (1986). Against parafoveal semantic preprocessing during eye
959 fixations in reading. *Canadian Journal of Psychology*, *40*(4), 473–483. https://doi.org/10.1037/h0080111
960 Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish
961 significance of effect latency or location. *Psychophysiology*, *56*(6), 1–8.
962 https://doi.org/10.1111/psyp.13335
963 Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*
964 *Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019
965 Serences, J. T. (2004). A comparison of methods for characterizing the event-related BOLD timeseries in rapid
966 fMRI. *NeuroImage*, *21*(4), 1690–1700. https://doi.org/10.1016/j.neuroimage.2003.12.021
967 Simola, J., Le Fevre, K., Torniainen, J., & Baccino, T. (2015). Affective processing in natural scene viewing:
968 Valence and arousal interactions in eye-fixation-related potentials. *NeuroImage*, *106*, 21–33.

https://doi.org/10.1016/j.neuroimage.2014.11.030

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. Non-linear effects, overlap correction, and practical considerations. *Neuropsychology*, *52*(2), 169–181. https://doi.org/10.1111/psyp.12320

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48. https://doi.org/10.1016/j.visres.2014.11.006

Thickbroom, G. W., Knezevič, W., Carroll, W. M., & Mastaglia, F. L. (1991). Saccade onset and offset lambda waves: relation to pattern movement visually evoked potentials. *Brain Research*, *551*(1–2), 150–156. https://doi.org/10.1016/0006-8993(91)90927-N

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology (2006)*, *59*(11), 1931–1949. https://doi.org/10.1080/17470210500416342

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, *17*(1), 159–170. https://doi.org/10.1016/j.concog.2006.11.008

Ušćumlić, M., & Blankertz, B. (2016). Active visual search in non-stationary scenes: coping with temporal variability and uncertainty. *Journal of Neural Engineering*, *13*(1), 1–12. https://doi.org/10.1088/1741-2560/13/1/016015

Viola, F. C., Debener, S., Thorne, J., & Schneider, T. R. (2010). Using ICA for the analysis of multi-channel EEG data. In *Simultaneous EEG and fMRI: Recording, Analysis, and Application: Recording, Analysis, and Application* (pp. 121–133).

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3), :24, 1-15. https://doi.org/10.1167/9.3.24

Võ, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: evidence from the

flash-preview moving-window paradigm. *Attention, Perception & Psychophysics*, *73*(6), 1742–1753. https://doi.org/10.3758/s13414-011-0150-6

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

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

Winkler, I., Debener, S., Müller, K.-R., & Tangermann, M. (2015). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2015-Novem*, 4101–4105. https://doi.org/10.1109/EMBC.2015.7319296

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, and Psychophysics*, *73*(6), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.00054

Yan, M., Richter, E. M., Shu, H., & Kliegl, R. (2009). Readers of Chinese extract semantic information from parafoveal words. *Psychonomic Bulletin and Review*, *16*(3), 561–566. https://doi.org/10.3758/PBR.16.3.561

1029 **Tables**

1030 Table 1

		Consistent Mean ± SD	Inconsistent Mean ± SD
Eye movement behaviour	Ordinal fixation number of first target fixation	7.7 ± 6.0	6.1 ± 5.4
	Fixation duration (*t*-2), in ms	220.7 ± 105	212.9 ± 95
	Fixation duration (*t*-1), in ms	207.6 ± 96	197 ± 91
	Fixation duration (*t*), in ms	261.6 ± 146	263.3 ± 136
	Gaze duration on target, in ms	408.5 ± 367.1	519.1 ± 373.6
	Number of re-fixations on target	1.7 ± 2	2.2 ± 2.1
	Duration of re-fixations on target, in ms	238.9 ± 121.8	250.2 ± 135.7
	Fixation duration (*t*+1), in ms	245.3 ±148	243.7 ± 146
	Incoming saccade amplitude to *t*-1 (°)	6.1 ± 5.2	5.9 ± 4.8
	Incoming saccade amplitude to *t* (°)	8.4 ± 5.2	8.2 ± 4.8
	Incoming saccade amplitude to *t*+1 (°)	9.5 ± 5.9	10.1 ± 5.8
	Distance of fixation *t*-1 from closest edge of target (°)	6.7 ± 9.92	6.3 ± 9.77
	Number of fixations after first encountering target object until end of trial	7.3 ± 2.1	7.2 ± 1.7
	Duration of fixations after first encountering target object (until end of trial)	254.6 ± 120.4	251.6 ± 118.8
Target object properties	Distance of target object centre from screen centre (°)	12.1 ± 2.8	11.7 ± 3
	Mean visual saliency (AWS model)	0.35 ± 0.16	0.37 ± 0.16
	Width (°)	6.1 ± 2	6.1 ± 2.1
	Height (°)	5.1 ± 1.8	5.4 ± 2.2
	Area (degrees of visual angle squared)	16.1 ± 8.7	17.3 ± 11.4

Note. Target object size and distance to target are based on the bounding box around the object. The fixation *t*+1 is the first fixation after leaving the bounding box of the target object.

Table 1. Eye movement behaviour in the task and properties of the target object.

Table 2

Predictor	Cumulative probability of First Fixation			
	β	*SE*	*z*-value	Pr(>|*z*|)
Intercept	-1.04	0.02	-49.91	0.00001
Nr. Fixation	-2.00	0.05	-35.9	0.00001
Consistency	0.17	0.03	5.8	0.00001
Consistency × Nr. Fixation	-0.71	0.09	-7.9	0.00001

Table 2. Cumulative probability of having fixated the critical object as a function of the ordinal number of fixations on the scene (binomial probit). The centred predictors are *Consistency* (Consistent: -0.5, Inconsistent: 0.5) and *Number of Fixation*

1047 Table 3

Predictor	Probability of immediate fixation			Latency to first fixation			Gaze Duration		
	β	*SE*	*z*	*β*	*SE*	*t*	*β*	*SE*	*t*
Intercept	-2.93	0.19	-14.73***	1,904.4	83.8	22.7***	400.1	20.97	19.08***
Consistency	0.21	0.15	1.38	-246.4	64.0	-3.85***	105.0	20.77	7.08***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1048 *Table 3*. Probability of immediate fixation and latency to first fixation. The simple coded
1049 predictor is *Consistency* (Consistent = -0.5, Inconsistent = 0.5). We report the *β*, the standard
1050 error, the *z*- (for binomial link) and *t*-value. Asterisks indicate significant predictors.
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066

Table 4

Type of Event	Analysis window	β	*SE*	*t*-value
Scene onset	Early (250-350 ms)	0.28	0.39	0.71
	Late (350-600 ms)	0.34	0.39	0.37
nt	Early (250-350 ms)	-0.06	0.07	-0.79
	Late (350-600 ms)	-0.09	0.08	-1.10
t -1	Early (250-350 ms)	-0.28	0.15	-1.77 (*)
	Late (350-600 ms)	-0.46	0.12	-3.76 ***
t	Early (250-350 ms)	-0.52	0.17	-3.03 **
	Late (350-600 ms)	-0.38	0.15	-2.43 *
t (control analysis with baseline before fixation *t*)	Early (250-350 ms)	-0.34	0.16	-2.20 *
	Late (350-600 ms)	-0.20	0.17	-1.14

(*) $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Mixed-effects models with maximal random structure for the FRP at the mid-central region-of-interest (comprising electrodes FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, and CP2) for two temporal windows of analysis (Early, 250-350 ms and Late, 350-600 ms) as predicted by *Consistency* (Consistent = -0.5, Inconsistent = 0.5) of which we report the β the standard error, and the *t*-value. Asterisks indicate the level of significance.

1104

1105 **Figures**

1106

Figure 1

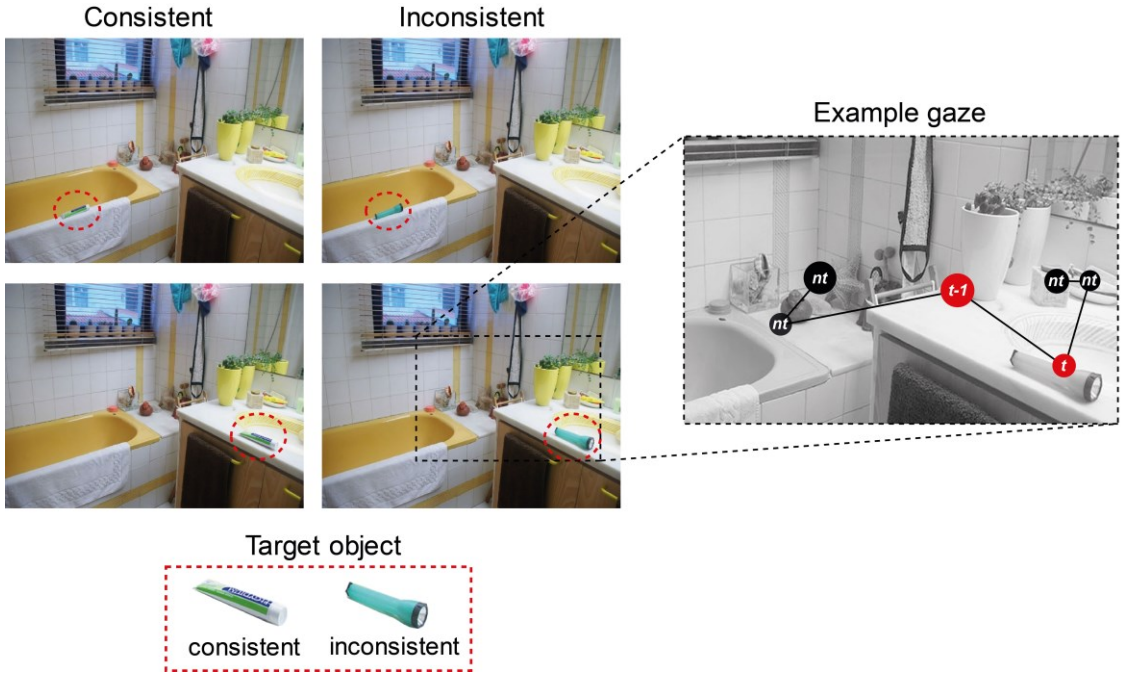

Figure 1. Example stimuli and conditions in the study. Participants viewed photographs of indoor scenes that contained a target object (highlighted with a red circle) that was either semantically consistent (here: toothpaste) or semantically inconsistent (here: flashlight) with the context of the scene. The target object could be placed at different locations within the scene, either on the left or the right side. The example gaze path plotted on the right illustrates the three type of fixations analysed in the study: (a) *t-1*; the fixation preceding the first fixation to the target object, (b) *t;* the first fixation to the target and (c) *nt;* all other (non-target) fixations. Fixation duration is proportional to the diameter of the circle, which is red for the critical fixations, and black for the non-target fixations.

1107

Figure 2

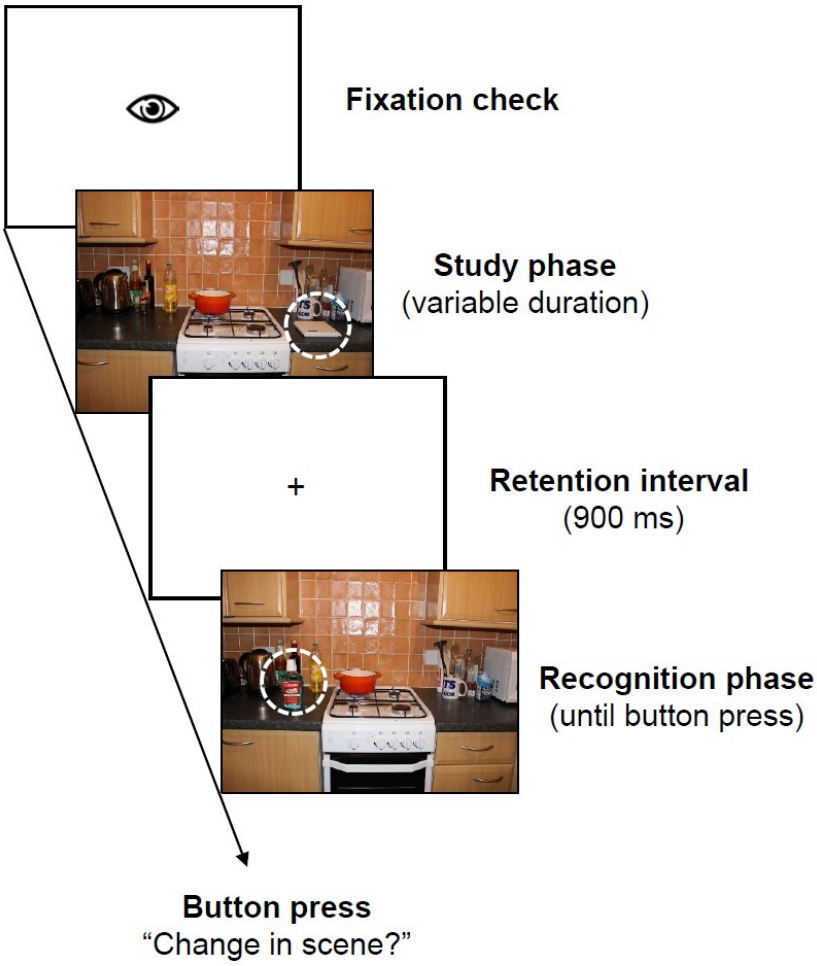

Figure 2. Trial scheme. Following a drift correction, the study scene appeared. The display duration of the scene was controlled by a gaze-contingent mechanism and it disappeared on average 2000 ms after the target object was fixated. In the following retention interval, only a fixation cross was presented. During the recognition phase, the scene was presented again until participants pressed a button to indicate whether or not a change had occurred within the scene. All analyses in the present paper focus on eye-movement and EEG data collected during the study phase.

1108

Figure 3

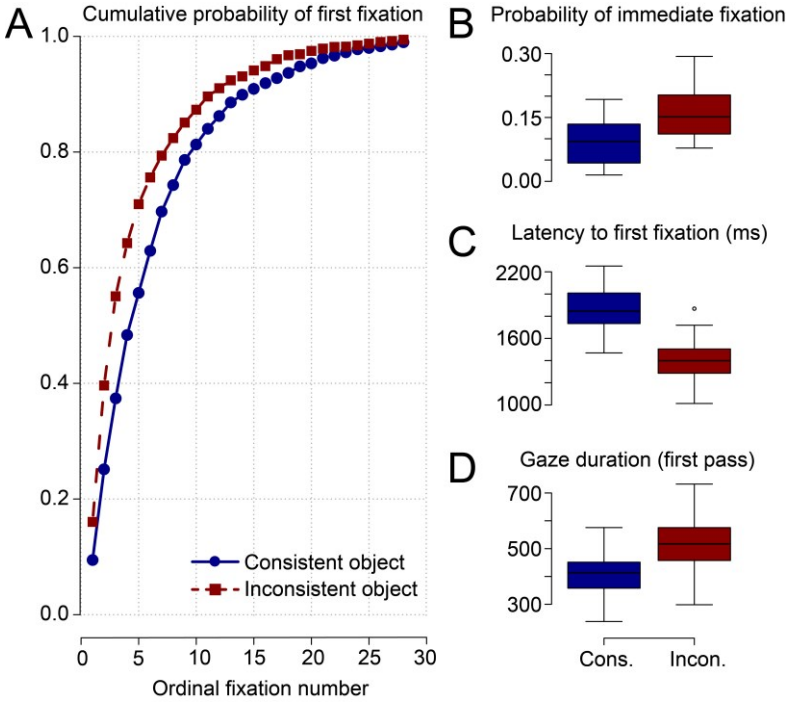

Figure 3. Eye-movement correlates of early overt attention towards consistent and inconsistent critical objects. **A.** Cumulative probability of fixating the critical object as a function of the ordinal fixation number on the scene. Blue solid line = consistent object; red dashed line = inconsistent object. **B.** Probability of fixating the critical object immediately, that is with the first fixation after scene onset. **C.** Latency until fixating the critical object for the first time. **D.** First-pass gaze duration for the critical object, i.e. the sum of all fixation durations from first entry to first exit. Whiskers of the boxplots (B, C, D) represent the 25th and 75th percentile of the measure (lower and upper quartiles). Dots indicates observations lying beyond the extremes of the whiskers.

1109

1110

Figure 4

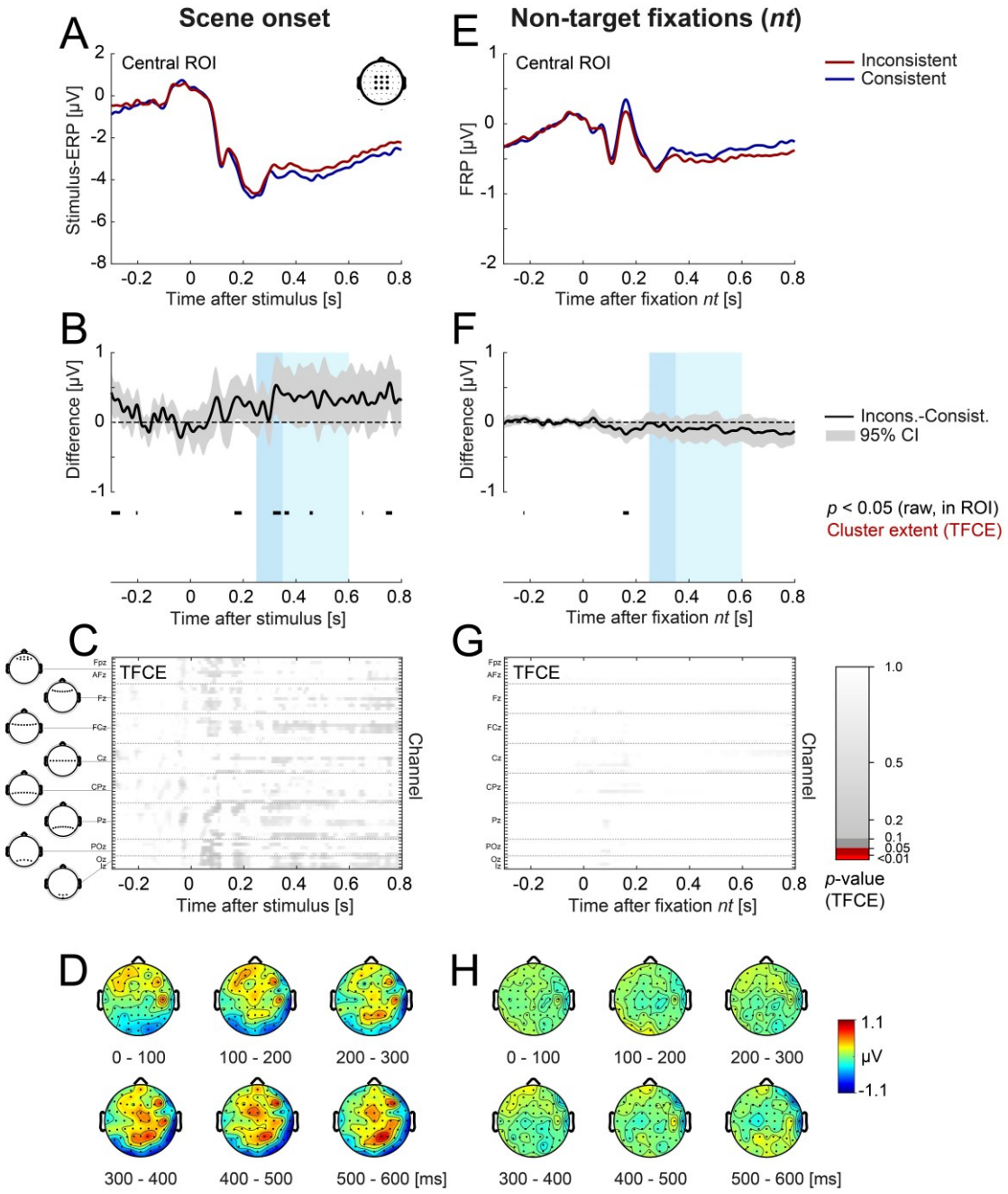

Figure 4. Stimulus-ERP aligned to scene onset (left panels) and FRP aligned to non-target fixations (right panels) as a function of object-scene consistency. **A and E.** Grand-average ERP/FRP at the central region-of-interest (comprising electrodes FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2). Red lines represent the *Inconsistent* condition, blue lines represent the *Consistent* condition. **B and F.** Corresponding difference waves (inconsistent minus consistent) at the central ROI. Grey shading illustrates the 95% confidence interval (without correction for multiple comparisons) of the difference wave with values outside the CI also marked in black below the curve. The two windows used for LMM statistics (250-350 and 350-600 ms) are indicated in light

blue. **C and G.** Extent of the spatiotemporal clusters underlying the cluster-based permutation statistic (TFCE) computed across all electrodes/time points. There were no significant ($p < 0.05$) effects. **D and H.** Scalp topographies of the consistency effect (inconsistent minus consistent) averaged across successive 100 ms time windows. Object-scene consistency had no significant effects on the stimulus-ERP or on the FRP elicited by non-target fixations, neither in the LMM statistic, nor in the cluster permutation test.

1111

Figure 5

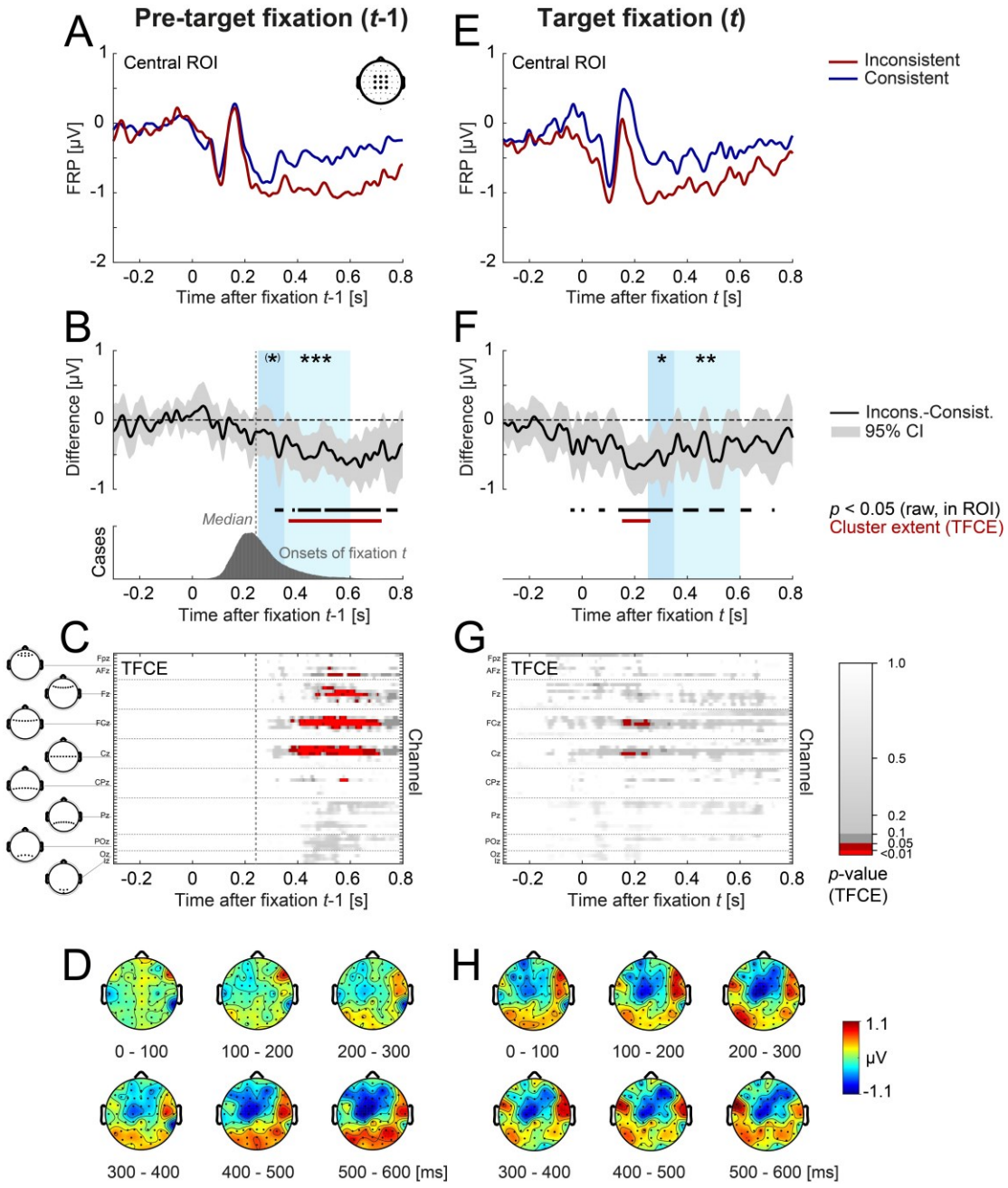

Figure 5. Grand-average FRP elicited by pre-target fixation (left panels) and target fixation (right panels) as a function of object-scene consistency. **(A, E)** Grand-average FRPs at the central ROI. **(B, F)** Difference waves at the central ROI. In panel B, the grey distribution shows the onset of fixation *t* relative to the onset of the pre-target fixation *t*-1, with the vertical dotted line indicating the mean latency (260 ms). **(C, G)** Results of cluster-based permutation testing (TFCE). The extent of the clusters from the first stage of the permutation test (marked in red) provides some indication which spatiotemporal features of the waveforms likely contributed to the overall significant effect of consistency. The temporal extent of the clusters is also illustrated by the red

bars in panels B and F. **(D, H)** Scalp topographies of the consistency effect (inconsistent minus consistent) across successive 100 ms time windows. A frontocentral N400 effect emerged in the FRP time-locked to fixation *t*-1 and reached significance shortly after the eyes had moved on to fixation *t*. This effect then continued during fixation *t* reaching a maximum 200 ms after the start of the target fixation.

1112

1113

Figure 6

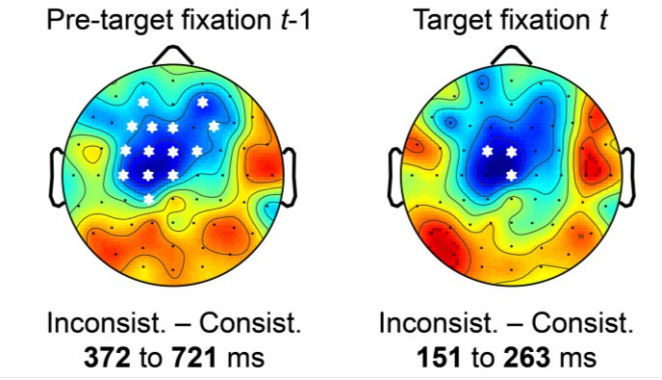

Figure 6. Scalp distribution of frontocentral N400 effects in the time windows significant in the TFCE statistic (see also Figure 5). White asterisks highlight the spatial extent of the clusters observed in the first stage of the TFCE permutation test for both intervals. In the FRP aligned to the pre-target fixation (left), clusters extended from 372 to 721 ms and across 13 frontocentral channels. In the FRP aligned to the target fixation (right), clusters extended from 151 and 263 ms at three frontocentral channels.

1114
1115