Can Prediction Error Explain Predictability Effects on the N1 during Picture-Word Verification?

Jack E. Taylor[1,2], Guillaume A. Rousselet[2], and Sara C. Sereno[2]

[1]Department of Psychology, Goethe University Frankfurt

[2]School of Psychology and Neuroscience, University of Glasgow

Author Note

Jack E. Taylor https://orcid.org/0000-0003-4765-0118

Guillaume A. Rousselet https://orcid.org/0000-0003-0006-8729

Sara C. Sereno https://orcid.org/0000-0001-7957-9542

Correspondence concerning this article should be addressed to Jack E. Taylor Department of Psychology, Goethe University Frankfurt; E-mail: Taylor@psych.uni-frankfurt.de

Abstract

Do early effects of predictability in visual word recognition reflect prediction error? Electrophysiological research investigating word processing has demonstrated predictability effects in the N1, or first negative component of the event-related potential (ERP). However, findings regarding the magnitude of effects and potential interactions of predictability with lexical variables have been inconsistent. Moreover, past studies have typically used categorical designs with relatively small samples and relied on by-participant analyses. Nevertheless, reports have generally shown that predicted words elicit less negative-going (i.e., lower amplitude) N1s, a pattern consistent with a simple predictive coding account. In our preregistered study, we tested this account via the interaction between prediction magnitude and certainty. A picture-word verification paradigm was implemented in which pictures were followed by tightly matched picture-congruent or picture-incongruent written nouns. The predictability of target (picture-congruent) nouns was manipulated continuously based on norms of association between a picture and its name. ERPs from 68 participants revealed a pattern of effects opposite to that expected under a simple predictive coding framework.

Keywords: N1, N170, Prediction, Predictive Coding, Word Recognition

Can Prediction Error Explain Predictability Effects on the N1 during Picture-Word Verification?

Introduction

Readers and listeners routinely use context to predict upcoming semantic and lexical content. Evidence for such predictive processes arises from both behavioural and neural correlates of language comprehension (Kuperberg & Jaeger, 2016; Luke & Christianson, 2016; Pickering & Gambi, 2018; Rayner et al., 2011; Van Petten & Luka, 2012), with demonstrated facilitation for the processing of predicted information (Federmeier, 2007; Pickering & Garrod, 2013).

A key question in this area is, how early in the processing stream are predictive processes able to modulate visual word recognition? One early stage in visual word recognition which may be sensitive to prediction involves the processing of visual word forms. A word form can be defined as the visual pattern of a single written word, comprised of smaller orthographic components (e.g., letters, letter bigrams, graphemes, strokes). While some electrophysiological evidence suggests sensitivity to orthographic variables in an earlier posterior P1 component peaking at around 100 ms after word presentation (e.g., Nobre et al., 1994; Segalowitz & Zheng, 2009; Sereno et al., 1998), the event-related potential (ERP) component most identified as an index of orthographic processing across different scripts is the first posterior negative-going wave, the N1 (Bentin et al., 1999; Lin et al., 2011; Ling et al., 2019; Maurer, Brandeis, et al., 2005; Maurer et al., 2008; Pleisch et al., 2019). The N1 is also sometimes referred to as the N170 due to the timing of its peak in some studies, at around 170 ms. This typically occipitotemporal, negative-going component shows reliable differences between orthographic and non-orthographic stimuli (e.g., words elicit more negative-going N1s than false-font strings do; Appelbaum et al., 2009; Bentin et al., 1999; Eberhard-Moscicka et al., 2016; Maurer, Brandeis, et al., 2005; Maurer, Brem, et al., 2005; Pleisch et al., 2019; Zhao et al., 2014).

Accounts of orthographic processing often stress the importance of top-down

predictions, and their interactions with bottom-up sensory input. For instance, the interactive account of the ventral occipito-temporal cortex (vOT), a region which is a likely generator of the N1 ERP component (Allison et al., 1994; Brem et al., 2009; Cohen et al., 2000; Dale et al., 2000; Maurer, Brem, et al., 2005; Nobre et al., 1994; Taha et al., 2013; Woolnough et al., 2021), suggests that sensitivity to orthography arises through the synthesis of bottom-up visuospatial information and top-down predictions informed by prior experience and knowledge (Price & Devlin, 2011). Such accounts exist within a predictive coding framework, according to which the brain utilises higher-level information to build, maintain, and continually update a generative model (or hierarchical series of generative models) of sensory information (Friston, 2010; Rao & Ballard, 1999; Rauss et al., 2011). A key feature of such accounts is that higher-level predictions cause lower-level features to be preactivated, and that the difference between the bottom-up sensory input and top-down predictions corresponds to a prediction error, which the brain attempts to minimise (Clark, 2013; Walsh et al., 2020).

In a predictive coding framework, prediction errors are determined by two key attributes: the magnitude of the error, and the precision or certainty of the error (Feldman & Friston, 2010; Kanai et al., 2015). Feldman and Friston (2010) likened the error signal to the calculation of the t statistic, where magnitude of an observation (i.e., mean, or mean difference) is divided by the inverse of its precision (i.e., standard error). Prediction errors, weighted by precision in this manner, can be conceptualised as representing the degree of "surprise" associated with a set of observations under a specified hypothesis.

Firstly, the magnitude of the error should determine the size of the error signal, with larger prediction errors resulting from greater mismatch between descending (top-down) predictions and ascending (bottom-up) sensory input. In neutral (non-biasing) contexts, a predictive coding account that includes learning of statistical regularities over extended periods would assert that error signals should vary as a function of stimulus regularity. More specifically, a predictive coding account of orthographic processing would expect error

signals to vary as a function of the size of the difference between a general orthographic prior (e.g., an average word form) and a presented word form. Some recent findings appear to support the notion that the N1 reflects a neutral-context error signal, with greater distance from an orthographic prior eliciting greater amplitude (Gagl et al., 2020), while the profile of the N1's sensitivity to word form regularity over experience matches that expected under a predictive coding account (Huang et al., 2022; Zhao et al., 2019).

Secondly, the precision or certainty of the prediction error should influence the response, with more certain descending predictions, and more certain ascending sensory input, eliciting greater error signals when predictions are violated. In neutral contexts, predictions, and certainty about them, may not be expected to vary much from a context-general prior. Indeed, it is easier to envisage the expected role of prediction precision for orthographic processing in biasing contexts, where precision is more variable than it is in neutral contexts. A predictive coding model of orthographic processing that allows for online, context-informed updating of orthographic priors would expect that the predictability of word forms should influence error responses, with more predictable contexts eliciting stronger error signals when word forms are prediction-incongruent, and weaker error signals when prediction-congruent. For instance, a sentential context that elicits a clear and reliable prediction for an upcoming word (i.e., that has high Cloze probability) should show a larger prediction error difference, between succeeding prediction-congruent and -incongruent word forms, than should a more neutral sentential context that is consistent with a large number of low-probability candidate words.

In this paper, we examine whether a simple predictive coding account that includes online updating of context-biased predictions and expectations can explain neural activity, captured in the N1, elicited by a word in context. Specifically, we examine whether sensitivity to prediction error in the N1 is dependent on contextual predictability, as a predictive coding account would expect. We hypothesise that according to a simple predictive coding model, the N1 should be larger for prediction-incongruent than

prediction-congruent word forms, in a manner dependent on the level of predictability, with greater differences at higher levels of predictability. Studies examining the effect of predictability on the N1's amplitude and latency have principally manipulated readers' expectations for specific visual word forms. Here, expectations are typically biased via linguistic contexts, where an initial text varies in how predictable it makes an upcoming target word form. In an alternative approach, bias is achieved via non-linguistic cues, such as cross-modal contexts and manipulation of task demands.

Biasing Word Form Predictions via Linguistic Cues

Readers' predictions of upcoming word forms are generally manipulated via linguistic cues. In these studies, a target word's predictability is typically determined in a pre-experiment norming study, operationalised via Cloze probability (i.e., the probability that the target is correctly guessed given its preceding context). Such a measure of word form predictability aligns closely with the concept of prediction precision or certainty in a predictive coding account.

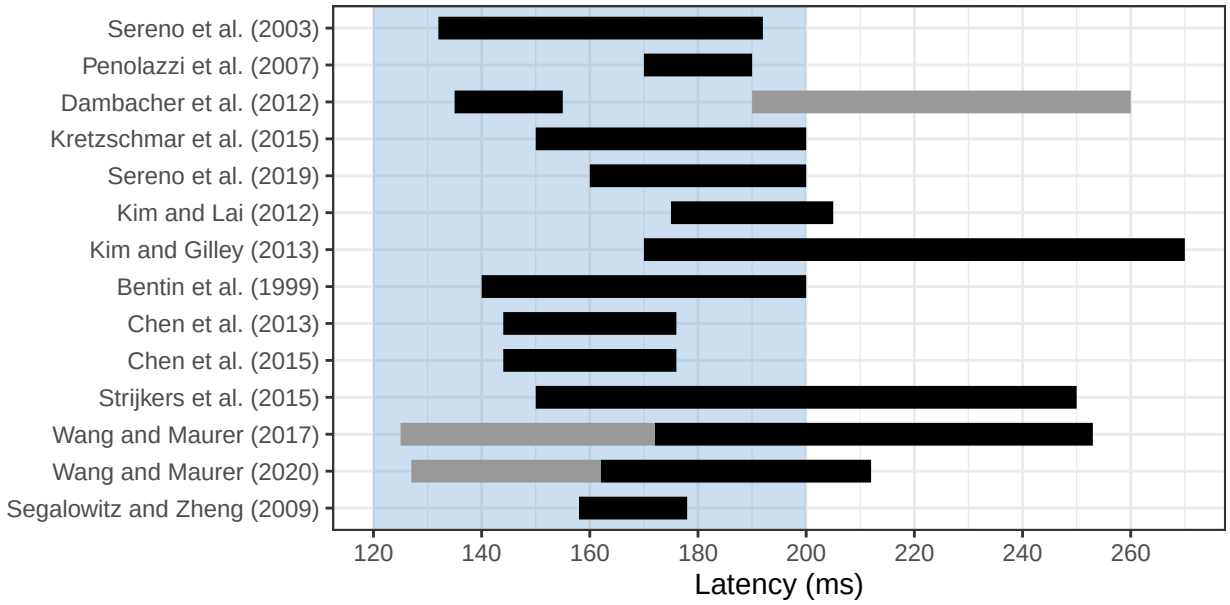
Recent ERP investigations that have manipulated sentential context have also often varied word frequency, with the assumption that an interaction of predictability with word frequency would provide evidence for top-down influences on lexical access. Such studies have demonstrated effects in the N1, although the pattern of effects observed across studies is varied (for a review, see Sereno et al., 2019). While effects often extend to earlier and later components, we limit our discussion to those involving predictability within the N1 window. Except where noted, electrodes analysed for the N1 were located occipitotemporally, and sentences were displayed word-by-word, using different word presentation rates or stimulus onset asynchronies (SOAs). It is important to note that the designated 'N1' time window differs across studies, as illustrated in Figure 1. Our review of studies utilising sentential contexts is presented chronologically. Sereno et al. (2003), using a 450 ms SOA, manipulated predictability (low, high) and word frequency (low, high), and found an interaction of these factors in the N1 (132-192 ms) across posterior and anterior

sites (comprising their first factor in a spatial factor analysis). Their predictability effect demonstrated less negative amplitudes at higher predictability, but only for low (and not high) frequency words. In a similar study, but using a 700 ms SOA, Penolazzi et al. (2007) manipulated predictability (low, high), word frequency (low, high), and, additionally, word length (monosyllabic words with an average of 4.08 letters, vs. disyllabic words with an average of 6.25 letters). In a 170-190 ms window, they found that high predictability conditions showed a more negative-going amplitude over centroparietal sites than low predictability conditions, but, unlike Sereno et al. (2003), found no interaction with word frequency. In addition, no significant interaction was observed with length. In a German study, Dambacher et al. (2012) varied predictability (low, high) and word frequency (low, high) in three experiments using different SOAs. At the shortest SOA of 280 ms, but not at SOAs of 490 or 700 ms, they found an interaction of predictability and frequency in the early portion of the N1 (135-155 ms). For high predictable words only, there was a frequency effect, with low frequency words showing a more negative-going amplitude than high frequency words over posterior sites. In a later N1 window, from 190 to 260 ms, Dambacher et al. reported an effect of frequency, but no interaction with (or main effect of) predictability. In a study measuring both eye movements and EEG during normal reading, Kretzschmar et al. (2015) manipulated items' predictability (low, high) and frequency (low, high). Testing only bilateral centroparietal electrodes, their fixation-related potentials (FRPs) demonstrated a main effect of predictability in a 150-200 ms window, with high predictable words showing a more positive-going amplitude than low predictable words, but without any interaction with frequency. Finally, Sereno et al. (2019) manipulated both predictability (low, high) and frequency (low, high). While the first, context sentence was presented in full, the second sentence containing the target word was presented word-by-word, with a short, 300 ms SOA. Sereno et al. (2019) found a predictability-frequency interaction in the N1 (160-200 ms). A predictability effect emerged only for high frequency words. Amplitudes to low predictable words, in comparison to those

to high predictable words, were more positive-going over left-hemisphere sites, but more negative-going over right-hemisphere sites. In sum, while these studies using sentential contexts have reported predictability effects in the N1 window, it is clear that the timing and topography of effects, as well as interactions with frequency, have been inconsistent.

Figure 1

N1 windows in predictability studies.



Some studies analysed two N1 windows (e.g., onset and offset). N1 windows reported to show a predictability effect are highlighted in black, while N1 windows that failed to show a predictability effect are highlighted in grey. Studies are listed in order of their mention in our review. For reference, the blue region displays the N1 period that we pre-registered.

Instead of manipulating error precision or certainty as the above studies have by varying predictability, A. Kim and Lai (2012) manipulated the orthographic error magnitude. Using a 550 ms SOA, the target word or alternative orthographic versions of it were presented in contexts that were acutely predictive of the target (M_{Cloze}=.90). Contexts were followed by the predictable target word (e.g., *cake*), an orthographically similar pseudoword (e.g., *ceke*), an orthographically dissimilar pseudoword (e.g., *tont*), or a consonant-string nonword (e.g., *srdt*). Consistent with an orthographic explanation for prediction effects in the N1, relative to targets, N1 (175-205 ms) amplitude was more

negative-going for orthographically dissimilar pseudowords and nonwords (i.e., when orthographic prediction error magnitude was greater). Orthographically similar pseudowords, while significantly different from all other conditions in the earlier P1, elicited N1 components more similar in amplitude to target words.

Another linguistic cue that has been manipulated is grammaticality. A. E. Kim and Gilley (2013) demonstrated effects of syntactic anomaly on the N1. Sentences leading to a strong prediction for the determiner, *the*, were presented unchanged or with the determiner replaced with an agrammatic preposition (e.g., *The thief was caught by the/for police*). The left-lateralised occipitotemporal N1 (170-270 ms) was more negative-going with the syntactically anomalous preposition than with the determiner. As the authors point out, the N1 effect is unlikely to be evidence for sensitivity to syntax per se. Rather, given evidence of the N1's sensitivity to orthographic features, it is probably more accurate to posit that the high predictability of the determiner's orthographic features elicited a less negative-going N1 when these predictions were confirmed.

A. E. Kim and Gilley (2013) simultaneous manipulation of orthography and syntax highlights a prevalent issue within the literature: namely, altering the visual word form necessitates alteration of the semantics, syntax, and/or plausibility of the sentence or wider discourse. Another limitation shared by studies using word-by-word presentation of sentences is that ERPs elicited by the target word can become difficult to disentangle from ERPs elicited by preceding or succeeding words, especially if the SOA is short or unjittered. While fast presentation times of sentential contexts and targets are useful for demonstrating that early modulation by predictive processes extends to realistic reading rates, their application may not be necessary to demonstrate that such modulation can occur. It is also of note that in a recent review of ERP studies using sentence- and discourse-level contexts to examine early neural correlates of word form prediction, Nieuwland (2019) concluded that findings thus far have been weak, inconsistent, and in need of more replication attempts. Moreover, most studies to date were not pre-registered

and often used inappropriate analysis models that did not account for measurement variability, raising questions about false positives in that literature.

Biasing Word Forms via Non-Linguistic Cues

Effects of prediction and expectation may alternatively be investigated using paradigms that modulate non-linguistic features of tasks and stimuli. In one approach, identical or suitably matched stimuli are presented under different task instructions (e.g., Compton et al., 1991). For instance, participants are more likely to predict and show sensitivity to lexical variables if given a word-nonword task than one which requires judgements on a non-lexical dimension, such as font colour. In a French study, Bentin et al. (1999) presented words, pseudowords, and consonant strings in a series of different tasks requiring participants to mentally count the word targets (oddballs) among word and nonword distractors. The tasks included lexical decision (word vs. nonword), semantic categorisation (i.e., abstract vs. concrete words or nonwords), and rhyme judgement (i.e., rhymes vs. does not rhyme with "-ail"). Although plots of ERPs and topographies suggest a trend towards a task-stimulus-hemisphere interaction on the N1 (140-200 ms), with the difference between orthographically plausible and implausible stimuli being larger in the tasks requiring lexical or semantic processing than in the rhyme task, the effect of task was not tested as a factor, such that any task-stimulus interaction is very difficult to interpret. Chen et al. (2013) compared ERP responses to target words in lexical decision and semantic categorisation (i.e., word vs. person's name) tasks to a condition with minimal task demands, namely silent word reading. They identified an effect of task on the N1 (144-176 ms), with a more negative-going N1 for words observed in lexical decision and semantic categorisation than in silent word reading. In a similar study, Chen et al. (2015) further suggested that the degree to which variables like frequency and imageability affect activity in the N1 (144-176 ms) was task-dependent. For instance, Chen et al. (2015) showed that increases in word frequency were associated with decreases in source-space activity during the N1, and that, crucially, this effect was larger in lexical decision than in

semantic decision (i.e., word vs. person's name) or silent word reading. In a French study examining word frequency effects across different go/no-go tasks, Strijkers et al. (2015) similarly reported that ERP amplitude in a period including the N1 (150-250 ms) was more sensitive to word frequency (with more negative amplitudes for higher frequency words) during a semantic categorisation (i.e., animal vs. non-animal) than a colour categorisation (i.e., blue vs. non-blue) task. Wang and Maurer (2017) applied a similar paradigm to examine how task modulated the effect of script familiarity on the N1 (125-253 ms) ERP. Chinese-reading participants were presented on each trial with either familiar Chinese characters or stroke-matched, unfamiliar Korean symbols, in three tasks: repetition detection, colour categorisation, and delayed naming (where participants would respond "symbol" to Korean characters). Wang and Maurer (2017) showed that the N1's sensitivity to character familiarity (with more negative amplitudes for unfamiliar Korean symbols than familiar Chinese characters) was greater in delayed naming and colour categorisation tasks than in a repetition detection task. This effect was specifically observed in the N1's offset period of 172-253 ms, where onsets and offsets are defined respectively as the periods in the component's time window which precede and succeed its peak. That the effect of character familiarity differed between colour categorisation and repetition detection is not straightforward to interpret, as these tasks may be expected to require similarly shallow processing of orthography. Nevertheless, the difference between delayed naming (which necessitates orthographic processing) and repetition detection (which does not) is suggestive of an effect of task demands on sensitivity to orthographic familiarity. Related non-sentential approaches to biasing participants' word form predictions include an attempt to alter expectations for different types of script. Wang and Maurer (2020) found that native Mandarin speakers' N1 sensitivity (onset 127-162 ms; offset 162-212 ms) to character familiarity, where unfamiliar Korean characters elicit greater N1 offset amplitudes than familiar Chinese characters, was greater when participants were led to expect Chinese characters.

In addition to task manipulations, non-sentential semantic contexts, leading to predictions for specific words or categories of words, have also been used to investigate predictive processing. In an ERP study, Segalowitz and Zheng (2009) presented words and pseudowords for lexical decisions in two conditions: words were either drawn from a single category (e.g., animals), or from five different semantic categories. Segalowitz and Zheng reported an interaction between stimulus type (word vs. pseudoword) and expectation (one vs. five categories) in the N1 (158-178 ms), wherein expectation affected N1 amplitudes for words but not for pseudowords. Their finding suggested that the N1 was sensitive to the greater predictive strength of a single semantic category. Using a similar paradigm, Hauk et al. (2012) compared ERPs in lexical (word vs. pseudoword) and semantic (living vs. non-living) decision tasks, showing that effects of category relevance were observed in the semantic decision task as early as 166 ms (data were analysed continuously, with no N1 window definition). This finding suggests, consistent with the findings of Segalowitz and Zheng, an early sensitivity to category relevance during the N1 which, given the N1's robust sensitivity to orthography, is likely to reflect an influence of semantic-level predictions on orthographic processing.

In another attempt to modulate top-down expectancy without linguistic context, Dikker and Pylkkanen (2011) implemented a picture-noun phrase verification task. An image of a target object alone or an image of objects related to the target object was followed by a written noun phrase (article + noun) denoting the target object. They manipulated congruency and predictability. For congruent trials, the noun phrase referred to a food/drink or animal (e.g., *the apple* or *the monkey*) that matched the prior image of the object presented on its own or 'contained' in a stylized image (e.g., a grocery bag or Noah's Ark, respectively). In the incongruent condition, the noun phrase did not match the prior image (single object or collection of objects). Predictability was considered high when the target object appeared on its own, and was considered low when the target object could be inferred to exist within the stylized images. Example conditions for the noun

phrase, *the apple*, are determined by its preceding image as follows: an apple (congruent, high predictability), a banana (incongruent, high predictability), a bag of groceries (congruent, low predictability), or Noah's Ark (incongruent, low predictability). Noun phrases (40 food/drink, 40 animal) were repeated four times across conditions. Although Dikker and Pylkkanen did not examine effects in the MEG equivalent of an N1 window, they did find effects of congruency only in the high predictive condition (i.e., the apple preceded by an apple vs. a banana image) in temporal windows preceding (∼100 ms) and succeeding (250-400 ms) the N1. Their stimuli were designed to minimise orthographic similarity between congruent and incongruent pairs of noun phrases (i.e., maximising the magnitude of orthographic errors), suggesting that the authors anticipated that any early sensory effect of predictability may be related to orthographic processing. With only 7 participants, the study likely lacked the sample size necessary to identify such an effect in an N1-like window. Indeed, in a related paradigm using fMRI, Kherif et al. (2011) presented picture and word prime-target pairs under four conditions: conceptual identity, semantically related, shared initial phoneme, and unrelated. In addition, the prime was either masked (after 33 ms) or not. The stimulus types of prime-target pairs were either matching (word-word with varying typography, or picture-picture with different views) or non-matching (word-picture or picture-word). In the unmasked conceptual identity condition, Kherif et al. showed priming effects in the left vOT (likely generator for the N1) for matching and non-matching stimulus types. Specifically, they showed that targets elicited reduced left vOT activity if preceded by congruent primes, regardless of whether the stimulus types were of matching or non-matching stimulus types. Assuming that picture identity is not directly processed in the left vOT, these findings suggest that higher-level processes link the identity and content of pictures to orthographic representations of word forms. However, Kherif et al.'s use of fMRI prevents interpretation of the timing of such effects - its coarse temporal resolution means that mapping of picture content to representations in vOT could occur so late as to be irrelevant to initial

orthographic word recognition processes.

One advantage of paradigms like picture-word verification tasks is that the researcher can control and manipulate variables like predictability and specificity of the picture-word relation. This was demonstrated in the design used by Dikker and Pylkkanen (2011), where the picture preceding the target word unambiguously biased participants' expectations to a single word form (with an image of one clearly identifiable object), or instead biased a set of semantically related possible word forms (with an image inducing multiple object candidates). Such a manipulation is comparable to the use of Cloze probability in sentential contexts or single versus multiple category priming, and similarly aligns with the concept of error precision or certainty.

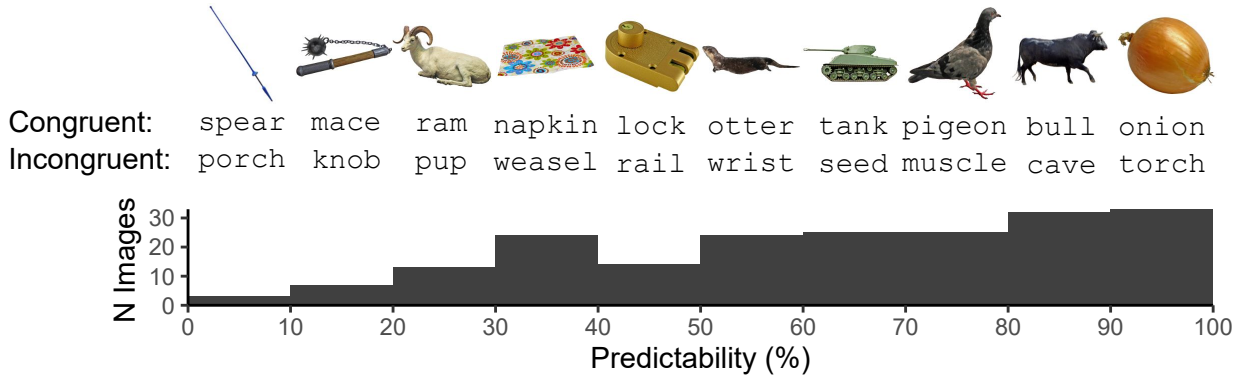
The Present Study

In the present study, we adapted the picture-word verification paradigm to examine the role of Predictability in prediction effects on the N1. We presented participants with PICTURE-word pairs that were congruent (e.g., ONION-onion) or incongruent (e.g., ONION-torch). Predictability of the congruent word was a continuous variable, dependent upon how often the noun is reliably used in naming the picture (**Figure 2**). By manipulating both Congruency and Predictability of word forms, we were able to examine whether the effect of Congruency on the N1 (sensitivity to prediction error) is contingent on Predictability (certainty or precision of prediction errors), in the manner expected according to a simple predictive coding account of the N1 in which observed N1 magnitude indexes prediction error.

We hypothesised, consistent with such a predictive coding account, that that there would be a Congruency-Predictability interaction in which at the highest levels of Predictability, N1s elicited by picture-incongruent words would be more negative-going than those elicited by picture-congruent words, while at the lowest level of Predictability picture-congruent and -incongruent words should elicit N1s of similar magnitude. We anticipated three patterns of results that would have been consistent with this hypothesis:

Figure 2

Illustration of the experimental stimuli.



PICTURE-word pairs were either congruent (e.g., NAPKIN-*napkin*) or incongruent (e.g., NAPKIN-*weasel*), while predictability of congruent picture-word pairs varied continuously. Ten example picture-congruent and -incongruent pairs are presented, with their predictability corresponding to the histogram bin they appear above.

(1) higher levels of Predictability lead to a reduction in N1 magnitude only for picture-congruent words, with no such effect for picture-incongruent words (**Figure 3a**);

(2) higher levels of Predictability lead to an increase in N1 magnitude only for picture-incongruent words, with no such effect for picture-congruent words (**Figure 3b**); or

(3) higher levels of Predictability lead to both a reduction in N1 magnitude for picture-congruent words and an increase in N1 magnitude for picture-incongruent words (**Figure 3c**).

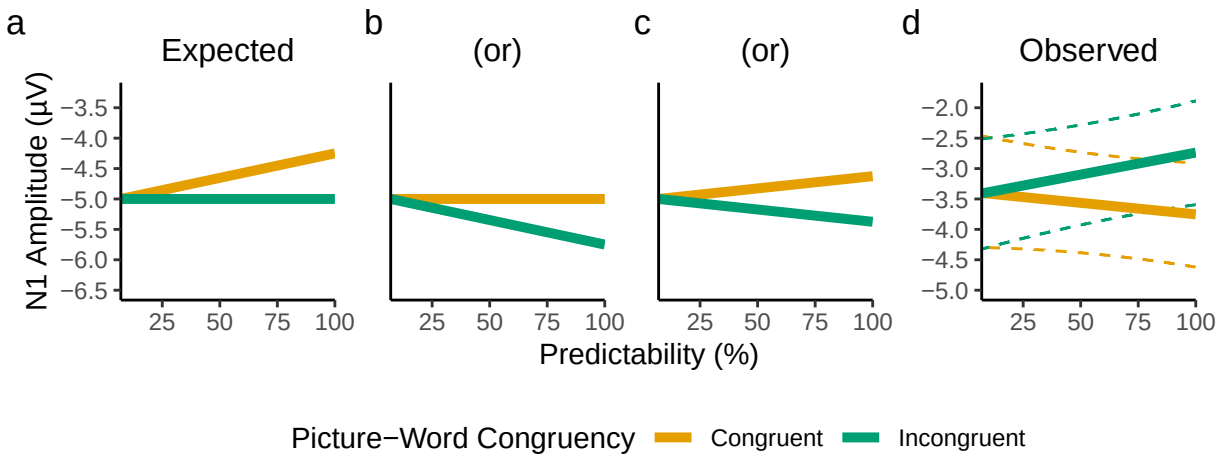
In our power analysis, we focused on the first of these possible patterns of results, but importantly, the Congruency-Predictability interaction term that we pre-registered to test our hypothesis (https://osf.io/jk3r4) would capture any of these patterns, as the interaction term's coefficient would be in the same direction in all cases.

In our analysis, we found a pattern of effects counter to our pre-registered hypothesis (**Figure 3d**), with a Congruency-Predictability interaction in the opposite direction. An exploratory Bayesian analysis revealed that the observed interaction was 59.98 times more likely than our hypothesis. Based on these findings, we argue our results suggest that such a simplistic predictive coding account is, at least on its own, insufficient to explain the

pattern of prediction effects observed in the N1 during a picture-word verification task.

Figure 3

*A comparison between the predicted (**a**,**b**,**c**) and observed (**d**) patterns of results.*



The predicted pattern of results was based on a predictive coding interpretation of the N1, according to which the magnitude of the N1 should be smaller for picture-congruent words relative to picture-incongruent words, and to a greater extent as Predictability increases. The observed pattern of results depicts the fixed effect predictions from the pre-registered linear mixed-effects model, with dashed lines depicting 95% bootstrapped prediction intervals (estimated from 5,000 bootstrap samples).

This study was pre-registered at https://osf.io/jk3r4 and the reported methodology and planned analysis conform to that specified in the pre-registration, except for two changes: an accidental change to timing of stimuli, and a lowering of the EEG high-pass filter cut-off. We explain these changes in the relevant sections, and demonstrate in **Supplementary Materials F** that the change to the high-pass filter cut-off had minimal effect on the results and conclusions. All data and code are available at https://osf.io/389ce/.

Method

The experiment included two separate tasks: The principal picture-word task was preceded by a localiser task to account for between-participant variability in the N1's timing and location. The details of stimulus selection and control as well as presentation

timing are provided in the following sections. For clarity, we first introduce the overall Congruency-Predictability design of the picture-word task. In this task, pictures of single objects are presented, followed by a noun, and participants decide whether the noun corresponds to the object. The level of Predictability of the noun was determined from norms of possible terms used to label a set of individual pictures (Brodeur et al., 2014). The most frequent, modal name agreement varied across pictures. Thus, level of noun Predictability was continuous and varied between 7% and 100%. The Congruency of the noun was either congruent (matching the modal name of the picture) or incongruent (a semantically unrelated noun matched across several lexical variables).

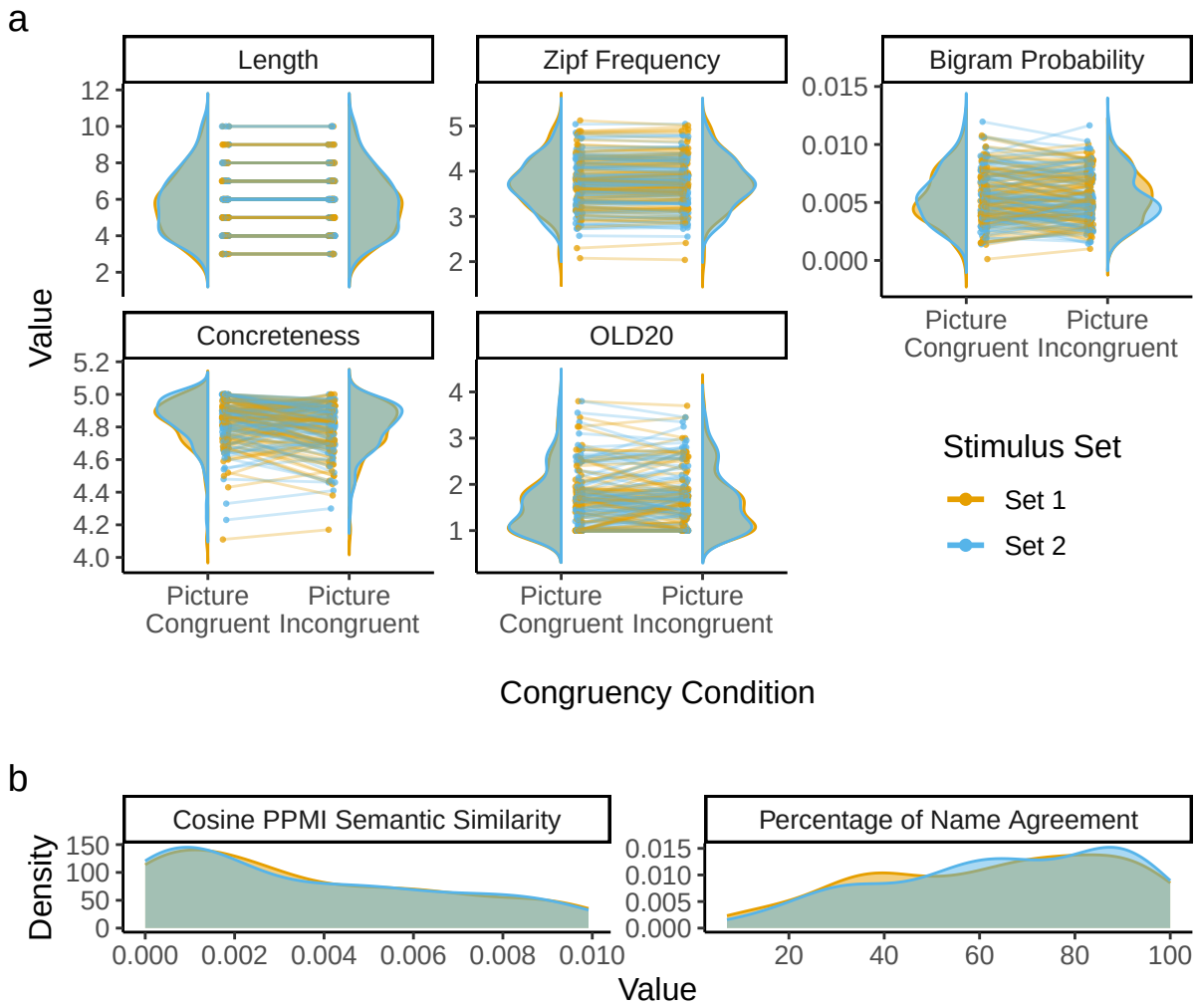
Materials: Picture-Word Task

A total of 400 words were selected with LexOPS (Taylor et al., 2020), a package for the generation and control of lexical variables in the R programming language (R Core Team, 2021). There were 200 words per Congruency condition, with one congruent and one incongruent word per image. A list of the full set of stimuli is available in **Supplementary Materials A**. The experimental stimuli are summarised in **Figure 4**. First, stimuli were filtered according to norms collected by Brysbaert et al. (2019), such that at least 90% of participants knew each word. In addition, stimuli were filtered such that all words were nouns according to the dominant part of speech data from SUBTLEX-UK (van Heuven et al., 2014), and had a mean concreteness rating above 4 (on a Likert scale from 1, least concrete, to 5, most concrete) according to Brysbaert et al. (2014). Images were taken from the Bank of Online Standardised Stimuli (BOSS) norms (Brodeur et al., 2014), a large database of images with normed statistics, including percentage of name agreement, which, critically, we used as a measure of Predictability. Words were identified as possible picture-congruent words if they were listed as the most frequent (i.e., modal) name for any image in the BOSS norms, and were identified as possible picture-incongruent words if they were not.

Picture-congruent and -incongruent words were matched item-wise across five

Figure 4

Summary of the picture-word stimuli.



Each panel depicts how a single variable was controlled. (**a**) Probability densities for variables which were matched item-wise between picture-congruent and picture-incongruent conditions, and distribution-wise between counterbalanced stimulus Sets 1 (in *yellow*) and 2 (in *blue*). Points representing pairs of words which are matched item-wise are joined by lines. Points' positions are jittered slightly along the x-axis for visibility. (**b**) Probability densities for two variables matched only in a distribution-wise manner between the counterbalanced stimulus sets: Cosine PPMI (Positive Pointwise Mutual Information) Semantic Similarity from SWOW (Small World of Words; De Deyne et al., 2019), and modal name agreement from the BOSS norms. These variables cannot be matched between Congruency conditions because only a single value describes each matched congruent-incongruent word pair.

lexical variables, with specific tolerance ranges, as follows: (1) word length (number of characters), exactly; (2) concreteness according to Brysbaert et al. (2014), within ±.25; (3)

Zipf frequency (a logarithmic scale of word frequency) according to SUBTLEX-UK, within ±.125; (4) character bigram probability (calculated from SUBTLEX-UK), within ±.0025; and (5) OLD20 (the average Orthographic Levenshtein Distance of the 20 closest neighbours to a given word; Yarkoni et al., 2008) calculated from the LexOPS inbuilt dataset, within ±.75. To ensure that picture-incongruent words were not inadvertent possible descriptors for images, the cosine positive pointwise mutual information (PPMI) measure of associative semantic similarity calculated from the Small World of Words (SWOW) word association norms (De Deyne et al., 2019) was minimised to be ≤.01 between each image's matched picture-congruent and picture-incongruent words. To ensure picture-incongruent words did not share orthographic features with their respective picture-congruent words, orthographic Levenshtein distance between matched items was maximised. As items were also matched in word length, this meant all matched pairs of words had a Levenshtein distance equal to their number of characters. The variable used to index the Predictability of picture-congruent words was percentage of modal name agreement, which was sampled pseudo-randomly (picture-congruent words were not selected if no incongruent match could be identified fitting the constraints specified above) from the BOSS norms, and varied continuously in the generated stimuli from 7 to 100%.

As the participants were recruited in the United Kingdom, possible congruent and incongruent picture-word pairs were excluded if we identified the words as less frequent in British English (e.g., *sidewalk*) or if they were modal names for images that the Canadian participants of the BOSS norms are likely to have been more able to name or distinguish (e.g., *buffalo*, *bison*). In addition, picture-word pairs were excluded if words were identified as shortened versions of nouns (e.g., *limo*, *chimp*) or alternate names for the same object (e.g., *motorbike*, *motorcycle*). Candidate picture-incongruent words were additionally excluded if images were not representative of the images in the BOSS (e.g., *waiter* or *church*, as there were no other images of people or entire buildings in the BOSS), or if they were unimageable despite their high concreteness value (e.g., *item*). Plural words (e.g.,

sticks) were excluded, as most images in the BOSS have modal names that are singular. Finally, four images with modal names *nut*, *trumpet*, *spinach*, and *tuba* were excluded, as we judged these names to be incorrect descriptions of their images.

To avoid repetition effects, each image was presented once, with participants viewing either the associated picture-congruent or picture-incongruent word. This was counterbalanced by splitting the stimuli pseudo-randomly into two equally sized stimulus sets, referred to as Set 1 and Set 2. Each participant was presented with only one of these stimulus sets. Pictures followed by congruent words in Set 1 were followed by incongruent words in Set 2, and vice versa. To minimise any systematic difference between the counterbalanced groups, the split of stimuli was selected to maximise the empirical distributional overlap (Pastore & Calcagnì, 2019) between the two stimulus sets in relevant variables. Specifically, the stimulus sets were selected from 50,000 random splits to maximise the overlap between the distributions of the following seven variables: (1) percentage of modal name agreement according to the BOSS norms; (2) cosine PPMI semantic similarity according to the SWOW; (3, 4) Zipf word frequency and character bigram probability according to SUBTLEX-UK; (5) word concreteness (Brysbaert et al., 2014); (6) word length; and (7) OLD20. Variables that were also matched item-wise between the conditions were matched distribution-wise separately within each Congruency condition. This ensured there were minimal systematic differences in distributions between conditions or stimulus sets.

To generate stimuli for practice trials, 20 matched pairs of picture-congruent and -incongruent words were generated using the same pipeline as above, except that word frequency, word concreteness, and character bigram probability were not matched item-wise. The practice stimuli were generated from images and words not used in the experimental stimuli. The same practice trials were presented to all participants.

Before embarking on the electrophysiological picture-word experiment, we first ran a proof-of-concept behavioural experiment using a different stimulus set generated from a

very similar pipeline. We anticipated that increased Predictability should cause faster response time (RT) for congruent trials and have either no effect or a minimal effect on performance for incongruent trials. The results from this behavioural validation are presented in **Supplementary Materials B**. In short, we observed the pattern of results consistent with our expectations, with Predictability leading to faster RTs for congruent trials, but having almost no effect on incongruent trials.

Materials: Localiser Task

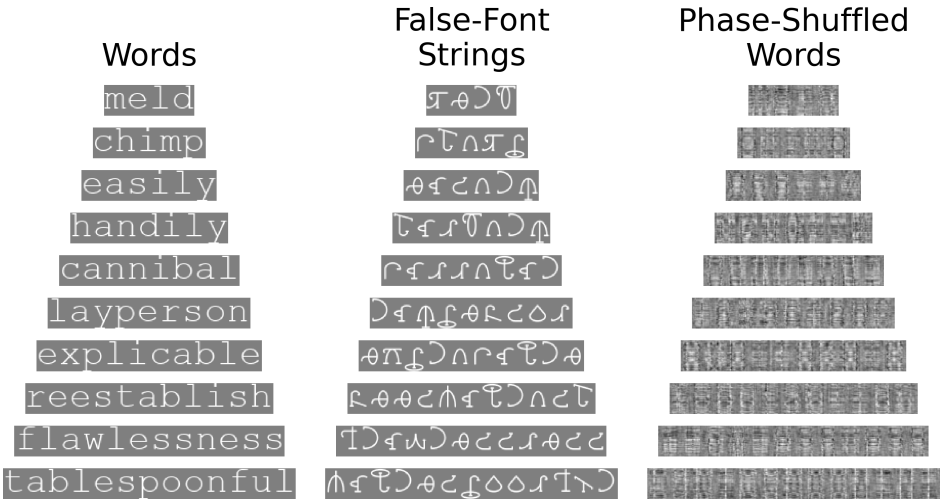
The precise location of the N1, and timing of its peak amplitude, is known to vary across studies and among participants. As such, we did not specify a common N1 electrode or timepoint shared among all participants before data collection. Instead, we employed a localiser task to identify, within an appropriate region and time period of interest, the electrode and timepoint at which each participant's maximal sensitivity to orthography emerges (i.e., more extreme amplitudes for words than false-font stimuli). This data could then be used to extract N1 amplitudes in the picture-word task, while accounting for variability among participants in timing and topography of orthographic processes.

For the localiser task, three categories of stimuli were presented for 100 trials each (**Figure 5**). These consisted of matched triplets of words (Courier New font), false-font strings (BACS2serif font), and phase-shuffled words. The comparison between words and false-font strings is a standard measure of N1 sensitivity to orthography, with previous evidence suggesting a more robust difference than exists between nonwords and words (Brem et al., 2018; Maurer, Brandeis, et al., 2005; Pleisch et al., 2019). However, phase-shuffled words were employed as an alternative comparison for exploratory analyses, with equal spatial-frequency amplitude and permuted spatial-frequency phase. Similar phase-shuffled word stimuli have shown robust differences to word forms in fMRI investigations of vOT activity (Rauschecker et al., 2012; Rodrigues et al., 2019; White et al., 2019; Yeatman et al., 2013).

To generate the localiser stimuli, a large list of suitable words (N=27,332) was

Figure 5

Ten example stimuli for each stimulus type in the localiser task.



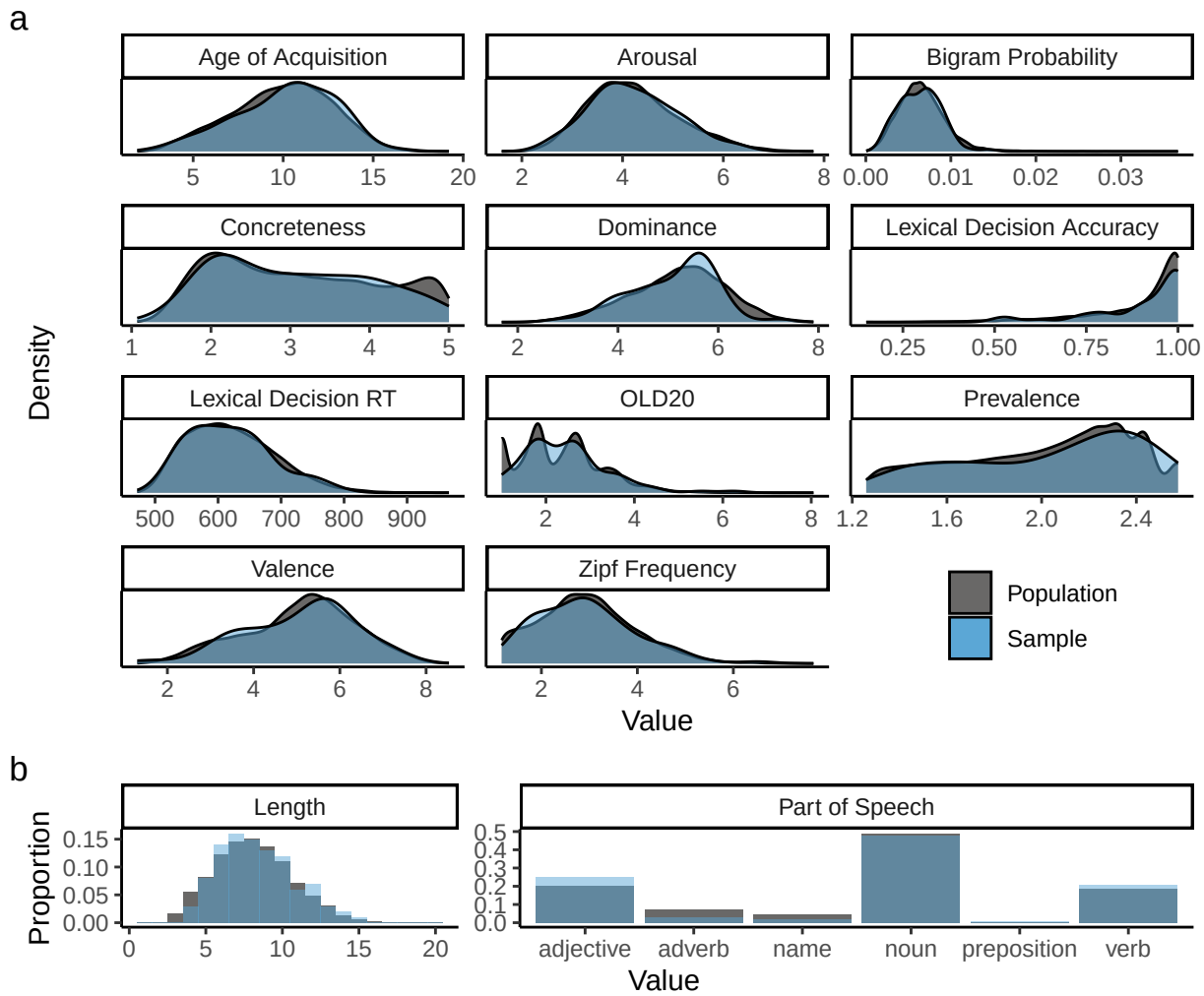
Each row represents a matched triplet of word, false-font string, and phase-shuffled word stimuli. The phase-shuffled word images were generated uniquely for each trial.

identified by filtering the word prevalence norms of Brysbaert et al. (2019) to only contain words known by at least 90% of participants and which were not selected for the main experiment. A representative sample (N=100) of this list was generated by maximising distributional overlap (Pastore & Calcagnì, 2019), between the sample and the full list of candidates, on 13 variables where observations were available: (1) word prevalence (Brysbaert et al., 2019); (2) length (number of characters); (3) word frequency in Zipf in SUBTLEX-UK (van Heuven et al., 2014); (4) part of speech according to SUBTLEX-UK; (5) character bigram probability calculated from SUBTLEX-UK; (6) OLD20 (Yarkoni et al., 2008) calculated from the LexOPS dataset (Taylor et al., 2020); (7) concreteness (Brysbaert et al., 2014); (8) age of acquisition (Kuperman et al., 2012); (9, 10) average lexical decision response time (RT) and accuracy according to the British Lexicon Project (Keuleers et al., 2012); and (11, 12, 13) the emotion ratings of valence, arousal, and dominance (Warriner et al., 2013). Similarity in the categorical variable of part of speech was maximised with dummy-coded variables (0 or 1 for absence or presence of a category, respectively). Distributional similarity across all variables was maximised by selecting from 500,000 random samples the sample with the highest total distributional overlap with the

full list of possible words. Distributions of the selected sample of words are summarised in **Figure 6**. The full list of stimuli for the localiser task is presented in **Supplementary Materials C**.

Figure 6

Distributions of key variables illustrate the similarity between the selected localiser stimuli words (sample) and the list of words from which they were drawn (population).



Panel **a** shows distributional similarity of continuous variables. Panel **b** shows similarity in length (all integer values) as a histogram showing proportions, and the similarity in the counts of each part of speech category as a bar plot of proportions. Only the part of speech categories which were present in the sample are shown. No members of less common part of speech categories, such as determiner or number, were selected in the sample.

The false-font strings consisted of characters from the Brussels Artificial Character

Set (BACS; Vidal et al., 2017) in BACS2serif font. In this way, we had an item-wise false-font match to each word, where every Courier New character in the word stimuli is replaced with a BACS character matched in the number of strokes, junctions, terminations, and serifs. The phase-shuffled stimuli were generated by using a Fourier transformation to extract the phase and amplitude from the word images. Phase values were randomly shuffled (i.e., permuted), such that the overall distribution of phase could be preserved, while amplitude values were unchanged. An inverse Fourier transformation was then used to generate a new image with the original amplitude values, but with phase randomly shuffled. To prevent phase shuffling from producing noticeably large changes in contrast, the phase shuffling was done on a version of the word image with 50% of the original contrast. After the inverse Fourier transformation, the contrast of the generated phase-shuffled image was readjusted to equal that of the original word image. To avoid repeating the same stimuli across participants more than necessary, unique phase-shuffled images were generated for each trial, for each participant.

Versions of the localiser task's stimuli were also created in green, to signal the participant to respond. For words and nonwords, this was done by simply changing the font colour to green. To preserve image intensity, the colour of phase-shuffled images was changed by altering pixels in the following way. For pixels in which the value in the green channel was less than 50% of the maximum intensity (i.e., the intensity of all channels in the grey background), values in red and blue channels were altered to equal the value in the green channel for that pixel. For all other pixels, the values in red and blue channels were set to 50% of the maximum intensity.

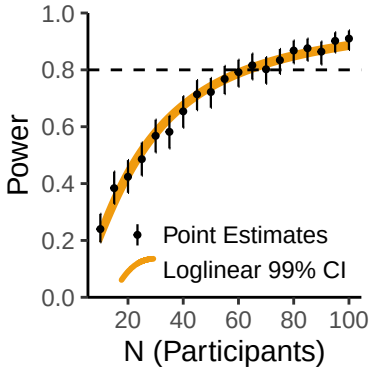
Participants

The sample size of 68 participants was decided via a power analysis using Monte-Carlo simulations of a realistic effect size (**Supplementary Materials D**). This revealed that with ≥68 participants we could expect >80% statistical power in the long run (**Figure 7**). All 68 participants (40 female, 27 male, 1 non-binary) were monolingual

native English speakers. Participants were randomly allocated into one of the four combinations of stimulus set (Set 1, Set 2) and response group (i.e., the left-right mapping of the two response buttons for affirmative and negative responses), such that each combination of stimulus set and response group comprised 17 participants. No participants reported diagnosis of any reading disorder. Ages ranged from 18 to 37 years (M=22.69, SD=4.9), and all participants reported having normal or corrected-to-normal vision. Participants' handedness was assessed via the revised short form of the Edinburgh Handedness Inventory (Veale, 2014), with participants only permitted to take part if they scored a laterality quotient of +40 indicating right handedness. Exclusion criteria for participants were determined prior to data collection as follows: (1) if 10 or more channels showed an offset more extreme than ±25 mV (as measured on the BioSemi acquisition software, ActiView), or (2) if more than 5% of the trials were lost due to technical issues with the EEG system. As no participants satisfied these criteria, no participants were excluded after data collection. Data collection was approved by the Ethics Committee of the institution at which the data were collected (application number: 300200117).

Figure 7

Estimated relationship between number of participants and statistical power.



Black points and error bars depict point estimates ±99% Binomial confidence intervals, each from 500 simulations. As 500 simulations provides a noisy estimate, we interpolated the relationship between N and power via a loglinear, logit-link Binomial model. The *orange* region depicts the 99% confidence intervals of this loglinear model.

Procedure

Stimuli were presented on a VPixx Technologies VIEWPixx screen (resolution 1920*1080 pixels, diagonal length 23", model VPX-VPX-2004A). Participants completed the experiment on a chin rest positioned 48 cm from the centre of the screen. Stimuli were presented on a grey background equal to 50% of the maximum intensity in each colour channel, roughly 12.3 cd/m^2. The experiment was written using the Python library PsychoPy (Peirce, 2007), and all code and materials are available in the repository associated with the study. All stimuli were presented centrally (horizontally and vertically). All trials in both tasks were presented in a pseudo-randomised order, such that no more than five consecutive trials required the same response from the participant. Trials were randomised across blocks, with the exception of the practice block, for which trials were randomised within the one block.

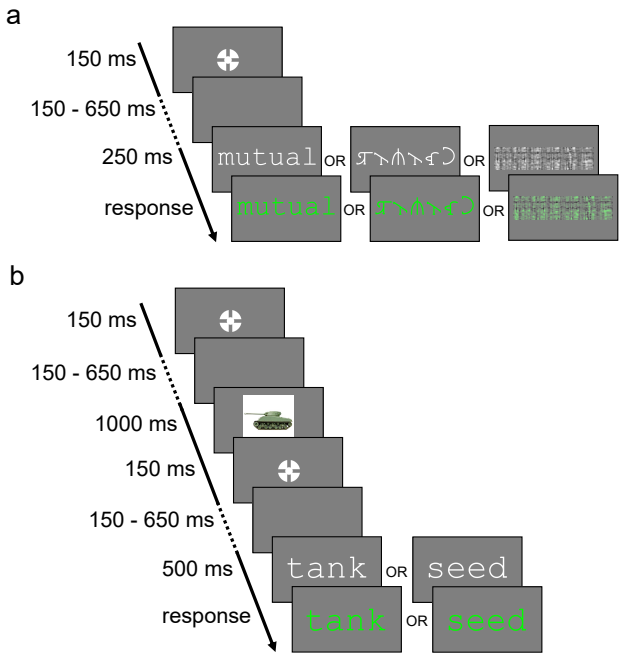
A mistake in the lab setup, which we discovered after data collection, meant that the display screen was running at 120 Hz rather than an expected 60 Hz. As we were controlling stimulus presentation by screen refreshes, this meant that all our stimuli were presented for half the expected durations. For this reason, the veridical stimulus durations described here differ from those described in the pre-registration.

Participants started with the localiser task, in the form of a lexical decision task (**Figure 8a**). The localiser task began with 30 practice trials, and was then followed by 300 trials split into 5 blocks of 60 trials. Each trial began with the bullseye fixation target recommended by Thaler et al. (2013) (outer and inner circle diameters were 0.6° and 0.2° of visual angle), presented for 150 ms. This was followed by a jittered interval of between 150 and 650 ms, during which the screen was blank. The stimulus (word, false-font string, or phase-shuffled word image) was then presented at a height of 1.5° (width of 1.07° for one character). Words and false-font strings were presented in white (80 cd/m^2), in the respective fonts of non-proportional Courier New and BACS2serif font. The stimulus was visible for 250 ms, after which the font colour changed to green to signal participants to

respond. Participants were requested to respond once after the stimulus changed colour, quickly and accurately, to indicate whether the stimulus they saw in each trial was either a word or not a word. The stimulus remained on screen until the participant responded. Responses were given with the right and left control ('Ctrl') keys of a QWERTY keyboard, with the mapping of affirmative and negative responses counterbalanced across participants. After the participant had responded, there was a delay of around 100 ms (variable as data was saved to disk during this interval), and then the next trial began.

Figure 8

*Trial structure of the (**a**) localiser task and (**b**) picture-word task.*



This figure is illustrative and the sizes are not to scale; in the experiment, images were in fact presented at a much larger scale than words.

After the localiser task, participants completed the picture-word task (**Figure 8b**), comprising an initial practice block of 20 trials, followed by 200 trials split into 5 blocks of 40 trials. As in the localiser task, each trial in the picture-word task began with the bullseye fixation point, presented for 150 ms, after which there was a blank screen for a jittered interval of between 150 and 650 ms. An image was then presented for 1000 ms, at a size of 10x10°. The bullseye fixation point was then presented again for 150 ms, followed

by another interval jittered between 150 and 650 ms. The word was then presented in white Courier New font, at a height of 1.5° (width 1.07° for one character). After 500 ms, the word turned green, and participants could provide their response to indicate whether the word described the image they saw. The word remained on screen until the participant responded. As in the localiser task, responses were given with the right and left control ('Ctrl') keys of a QWERTY keyboard, with the mapping of affirmative and negative responses counterbalanced across participants, but kept consistent within participants across the two tasks. After participants had responded, there was a delay of around 100 ms (again, variable as data was saved to disk during this interval), and then the next trial began. There was no deadline for participants to respond. The instructions given to participants for the picture-word task are presented in **Supplementary Materials E**.

The first blocks of both tasks consisted of practice trials with 10 exemplars for each stimulus type (word or false-font string or phase-shifted image, and congruent or incongruent noun for the localiser and picture-word tasks, respectively), during which participants were additionally given immediate feedback on their accuracy for each trial. These practice trials were followed by green text reading "CORRECT!" if the participant responded correctly, or else by red text reading "INCORRECT!", presented in Courier New font with a height of 1.5°, for 1000 ms. Participants had self-paced breaks between blocks for each task. Before the practice trials and at the start of every experimental block, participants were presented with instructions for the task (available in **Supplementary Materials E**), summarising what would occur in each trial, and specifying that they should respond as quickly and accurately as possible once the stimulus turned green. These instructions also specified which keys participants should press to indicate their decision. After each experimental block, including the practice trials, participants were presented with their average accuracy and median response time. After the practice trials, participants were additionally given the option to run the practice trials again. In the experimental blocks, no trial-level feedback was provided.

Recording

EEG data were recorded using a 64-channel BioSemi system, sampling at 512 Hz, with an online low-pass filter at the Nyquist frequency. Electrodes were positioned in the standard 10-20 system locations. Four electro-oculography (EOG) electrodes were placed to record eye movements and blinks: 2 were placed to the sides of eyes (on the right and left outer canthi), and 2 below the eyes (on the infraorbital foramen). Electrode offset was kept stable and low through the recording, within ±25 mV, as measured by the BioSemi ActiView EEG acquisition tool. Electrodes whose activity exceeded this threshold were recorded but were removed (and interpolated) in data preprocessing.

Preprocessing

The following section details the procedure applied to EEG data from each individual session, with the same pipeline being applied to both the localisation task and picture-word task unless otherwise specified. EEG preprocessing was achieved using functions from the EEGLAB (Delorme & Makeig, 2004) toolbox for MATLAB (MATLAB, 2022) or OCTAVE (Eaton et al., 2020). For both tasks, trials were excluded if responded to incorrectly (N=368, or .02%, in localiser task, N=226, or .02%, in picture-word). Further trials were excluded if responded to later than 1500 ms after the word (or nonword) changed colour (N=41, or .002%, in localiser task, N=42, or .003%, in picture-word).

Channels recorded as having offsets ±25 mV during data acquisition were removed from the data (in both tasks, 56 channels, or 1.27%, were removed across all participants), with their activity to be later interpolated. The EEG data were then re-referenced to the average activity across all electrodes and filtered with a 4th order Butterworth filter between .1 and 40 Hz. To counteract the distortion in signals' timing (phase) that is inherent to causal filters, the filter was applied in both directions (i.e., two-pass), with the MATLAB function, *filtfilt()*. In our pre-registration, we specified that we would apply a Butterworth filter with a bandpass of .5-40 Hz. However, after the pre-registration, we considered that, consistent with research into the effects of high-pass filters (Rousselet,

2012; Tanner et al., 2015; VanRullen, 2011), this could produce artefactually early effects. As a result, we lowered the high-pass filter to a less problematic .1 Hz. For comparison, demonstrating that our change to the pre-registered pipeline had minimal effect on the results or our conclusions, the results using the original filter are presented in **Supplementary Materials F**.

Segments of data outside of experimental blocks (i.e., in break periods) were identified and removed so they did not impact the independent components analysis (ICA) applied later in the pipeline. Blocks were identified as beginning 500 ms before stimulus presentation in the first trial of each block, ending 500 ms after the end of the last trial's epoch. To reduce the impact of occasional non-stationary artefacts with high amplitude (such as infrequent muscle movements), artefact subspace reconstruction (ASR; Chang et al., 2020) was used with a standard deviation cutoff of 20 to remove non-stationary artefacts. Following this, an ICA was run on the data to identify more stationary artefacts. The ICA was run using the FastICA algorithm (Hyvärinen & Oja, 1997), with a recorded random seed for reproducibility. The ICA was run on a copy of the data with channel offsets removed to allow for better sensitivity to electro-oculogram (EOG) artefacts (Groppe et al., 2009). The ICLabel classifier (Pion-Tonachini et al., 2019) was used to automatically identify artefacts which were eye- or muscle-related. Components classified by ICLabel as eye-related or muscle-related with a probability of $\leq$85% were removed from the data. Following eye movement artefact removal, activity from channels which were removed was interpolated via spherical splines (Localiser: M=1.14 per participant, SD=1.58; Picture-Word: M=1.68, SD=2.03), as implemented in EEGLAB. Trials were then epoched and baseline-corrected to the 200 ms preceding stimulus presentation. For the localiser task, stimulus presentation refers to the time point at which words, false-font strings, or phase-shuffled images were presented; in the picture-word task, stimulus presentation refers to the target word.

For the planned analysis, we pre-registered an approach to maximise sensitivity to

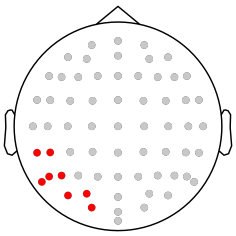
effects of Congruency and Predictability on the N1. To encompass the typical topography and timing of the posterior left-lateralised N1, we selected eight occipitotemporal electrodes (**Figure 9**; electrodes O1, PO3, PO7, P5, P7, P9, CP5, and TP7) and a 120-200 ms window. In contrast to some previous studies whose N1 windows extended beyond 200 ms, we set 200 ms as an upper bound for the possible maximal timepoint in the main analysis, to ensure effects were indeed restricted to the N1, and not later components like the N400. For each participant, we identified the electrode that showed maximal sensitivity to orthographic information in the N1 during the localisation task. Specifically, each participant's "maximal electrode" (within the region of interest and selected time window) was the one which showed the largest mean amplitude difference, in the expected direction, across all localiser trials between word and false-font string stimuli. The expected direction was a more negative-going N1 for words than for false-font strings, a pattern based on previous findings (Appelbaum et al., 2009; Bentin et al., 1999; Eberhard-Moscicka et al., 2016; Pleisch et al., 2019; Zhao et al., 2014). Each participant's "maximal timepoint" was the timepoint at which the maximal electrode showed the greatest sensitivity to the word-versus-false-font difference in the expected direction. Each participant's maximal electrode and maximal timepoint were then used to extract their trial-level N1 amplitudes from the picture-word task. To reduce the influence of noise on trial-level data, the trial-level N1 amplitudes in the picture-word task were calculated as the maximal electrode's mean amplitude across 3 timepoints: the participant's maximal timepoint, and the timepoints immediately preceding and following it. At the recorded sample rate of 512 Hz, this is equivalent to a window of 5.85 ms (i.e., 1/512*3) centred on the maximal timepoint.

Results

The planned analysis (pre-registered at https://osf.io/jk3r4) examined whether the hypothesised effect of a Predictability-dependent reduction of N1 amplitudes for picture-congruent words was observed at the electrode/timepoint in which each participant

Figure 9

The left-lateralised occipitotemporal region of interest selected for the N1 (highlighted in red).



showed maximal sensitivity to orthography. We then present exploratory analyses, which respectively examine the Bayesian probability that our data are consistent with the hypothesis, and delineate the time-course of the Congruency-Predictability interaction. We also conducted exploratory behavioural analyses, which we report in the supplementary materials, examining behavioural results in the picture-word study (**Supplementary Materials G**), and EEG and behavioural results from the localiser task (**Supplementary Materials H**)

Planned Analysis

The planned analysis tested the pre-registered hypothesis of a Congruency-Predictability interaction in which N1 amplitudes are reduced (i.e., less negative going) for picture-congruent trials than for picture-incongruent trials, and in which this difference is greatest at the highest levels of predictability, and smallest at the lowest levels of predictability. This was based on the notion that the N1 indexes prediction error in biasing contexts. We did not find evidence in favour of this hypothesis.

The trial-level N1 amplitudes from the picture-word task were modelled using a linear mixed-effects model fit with the R package *lme4* (Bates et al., 2015), estimating the maximal random effects structure justified by the experiment's design (Barr et al., 2013) as detailed in the section on the power analysis. The model was fit using the *bobyqa* optimiser (Powell, 2009). In *lme4* syntax, the formula for the mixed-effect model was specified as:

```
amplitude ~ 1 + congruency * predictability +
```

```
    (1 + congruency * predictability | participant_id) +
    (1 + congruency | image_id) +
    (1 | word_id)
```

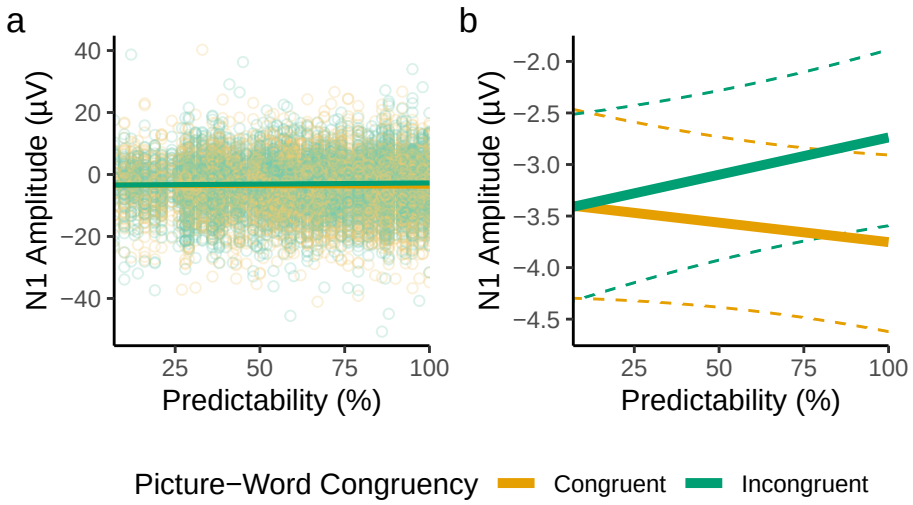
In this formula, *amplitude* is the trial-level N1 amplitude in microvolts, while *congruency* is a deviation-coded categorical variable indicating whether a given trial's word was picture-congruent or -incongruent, and *predictability* refers to the proportion of name agreement in the BOSS norms, normalised between 0 and 1. A consequence of this coding method is that the model's intercept reflects the predicted amplitude at the lowest level of Predictability, averaged across both levels of Congruency, while the slopes' coefficients are standardised and directly comparable in their magnitude. The variables of *participant_id*, *image_id*, and *word_id*, in the formula, identify each trial's participant, image, and word, respectively.

The fixed effect relationships predicted by the model are presented in **Figure 10**. The model intercept, reflecting the average N1 amplitude at the lowest level of Predictability, was estimated to be β=-3.35 µV (SE=.50). The fixed effect of Congruency from this model was estimated as β=-.02 µV (SE=.32), which captures that, at the lowest level of Predictability (7%), N1 components for picture-congruent and -incongruent words were estimated to be very similar (.02 µV difference). The main effect of Predictability was estimated as β=.16 µV (SE=.29), meaning that N1 amplitudes, averaged across congruent and incongruent conditions, were only .16 µV less negative-going at the highest level (100%) than at the lowest level of Predictability (7%). The effect of interest, the interaction between Congruency and Predictability, was in the opposite direction from that hypothesised, estimated as β=-1.02 µV (SE=.50). As the polarity of our predictions was explicitly specified, we interpret these results as a failure to find evidence in favour of the hypothesis.

To describe the estimated interaction, for picture-incongruent words, the effect of Predictability was estimated to be β=.66 µV (SE=.35), while for picture-congruent words,

Figure 10

Fixed effect predictions from the planned analysis of the picture-word task.



(**a**) Model-derived fixed-effect predictions, visualised over results from all trials (individual points). (**b**) Fixed-effect predictions visualised alone for visibility, with dashed lines depicting the bounds of 95% bootstrapped prediction intervals (estimated from 5,000 samples), where bootstrapped predictions were generated using the *bootMer()* function of *lme4*. For feasibility, bootstrapped predictions were generated from a version of the model that lacked random slopes.

the effect of Predictability was estimated to be β=-.37 µV (SE=.38). As such, the slopes for the effect of Predictability in both Congruency conditions were in directions inconsistent with our predictive coding hypothesis.

For comparison, we also analysed the data altering aspects of our planned analysis method: first using the maximal electrodes that would be identified from the comparison between words and phase-shuffled words, and second using averages within the occipitotemporal region of interest (**Supplementary Materials I**). These exploratory analyses revealed very similar patterns of effects, with estimates of the Congruency-Predictability interaction similarly inconsistent with our hypothesis, which we derived from a simple predictive coding account of the N1.

Exploratory Bayesian Analysis

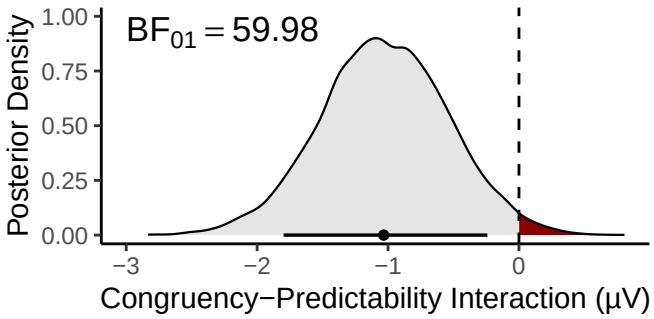
We observed a Congruency-Predictability interaction in the opposite direction (i.e., negative) to what we expected under our predictive coding hypothesis (i.e., positive). To explicitly quantify the probability of our predictive coding hypothesis, we fit a Bayesian implementation of the model described in the planned analysis, in STAN (STAN Development Team, 2023) via *brms* (Bürkner, 2017). This model was fit to the same data, and estimated the same hierarchical formula, with the same Gaussian link function as that described above, but was specified with weakly informative priors for the fixed effects. Specifically, the prior for the fixed effect intercept was specified as a normal distribution of mean -5, and *SD* 10, while all fixed effect slopes' priors were specified as normal distributions centred on 0, with *SD*s of 5. Covariance matrices were assigned flat priors, and default priors for brms were used for random effect *SD*s and the sigma parameter of the normal distribution. The model was fit with 5 chains and 5000 iterations per chain (split equally between warmup and sampling) such that there were a total of 12,500 posterior samples. Consistent with the linear mixed-effects model we fit via *lme4*, this analysis revealed a median posterior estimate for the Congruency-Predictability interaction of β=-1.03 µV (89% highest density interval = [-1.8, –.24]; **Figure 11**). We calculated, given this posterior distribution, that the Congruency-Predictability interaction is 59.98 times more likely to be less than 0, than it is to be greater than zero (that is, BF_{01}), which we consider to be strong evidence against our hypothesis.

Exploratory Time-Course Analysis

To examine the time-course of effects, we fit separate linear mixed-effects models to sample level data for the left-lateralised occipitotemporal region of interest, with variables coded as described for the planned analysis. For feasibility, data were downsampled to 256 Hz, and the models did not estimate random slopes. To account for variability between electrodes, and for per-participant differences in topography, random intercepts were estimated for each combination of participant and electrode. In *lme4* syntax, the model

Figure 11

Posterior density for the Congruency-Predictability interaction.



The region of the posterior distribution consistent with the predictive coding hypothesis (where $\beta>0$) is highlighted in *red*. The point and horizontal line below the density plot depict respectively the median estimate and 89% highest density interval of the posterior distribution.

formula was specified as follows:

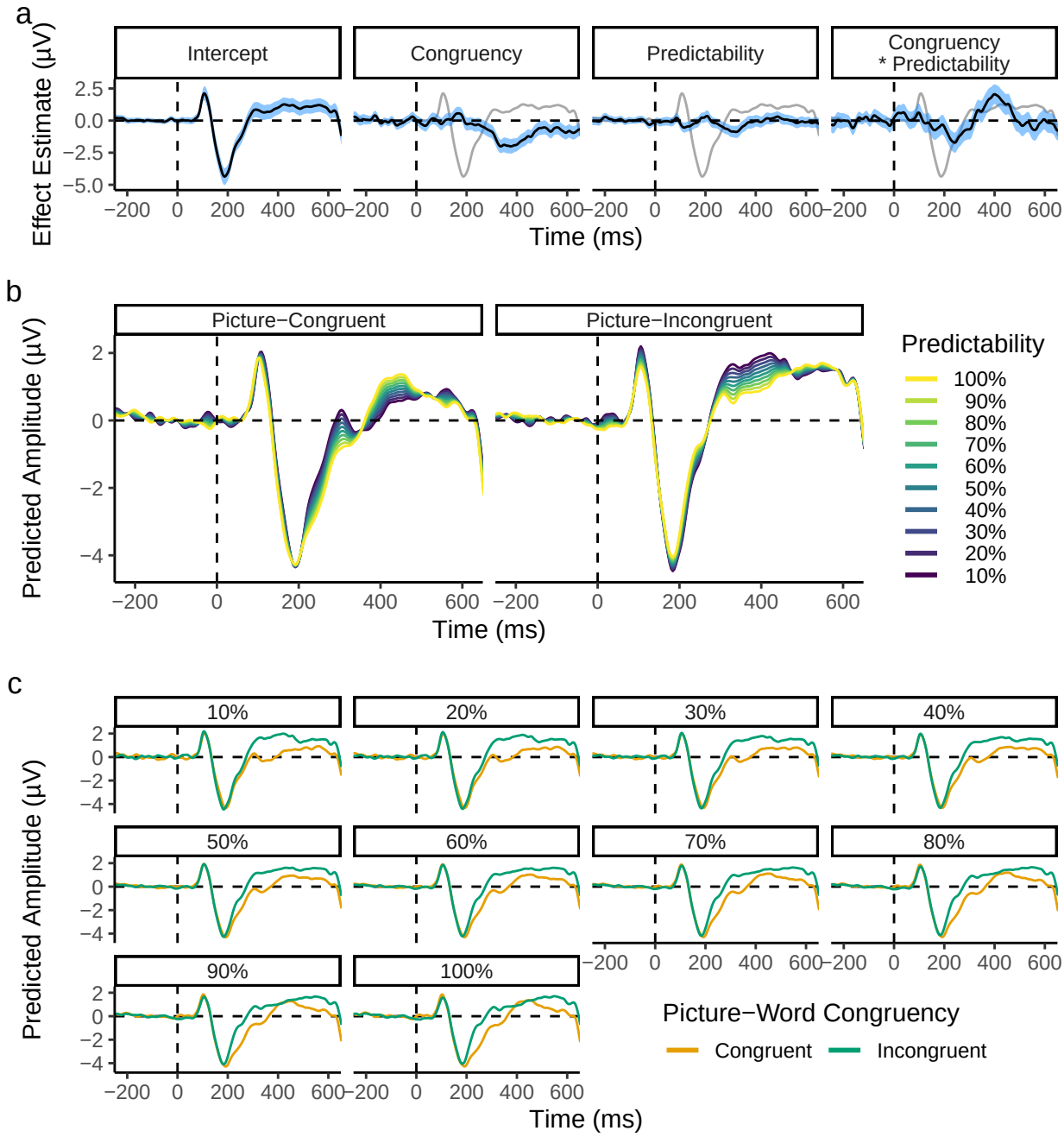
```
amplitude ~ 1 + congruency * predictability +
    (1 | participant_id) +
    (1 | participant_id:electrode_id) +
    (1 | image_id) +
    (1 | word_id)
```

The results (**Figure 12**) reproduced findings from the planned analysis, with increases in Predictability associated with more negative (larger) N1 amplitudes for picture-congruent words, and with less negative (smaller) N1 amplitudes for picture-incongruent words. The Congruency-Predictability interaction of interest remained negative, and thus in the opposite direction to that hypothesised, throughout the N1.

The sample-level analysis additionally suggested that the difference was largest in the N1's offset period (succeeding the peak). A later Congruency-Predictability interaction was also observed, peaking at around 400 ms (possibly resulting from effects in the N400 component) in the opposite direction to that observed for the N1's offset. To better understand the time-course of the Congruency-Predictability interaction, we examined the time-course of the effect of Predictability for picture-congruent and -incongruent words separately (i.e., simple effects; **Figure 13**). This showed more clearly that Predictability

Figure 12

Time-course of fixed effects from the sample-level analysis of the left-lateralised occipitotemporal region of interest.

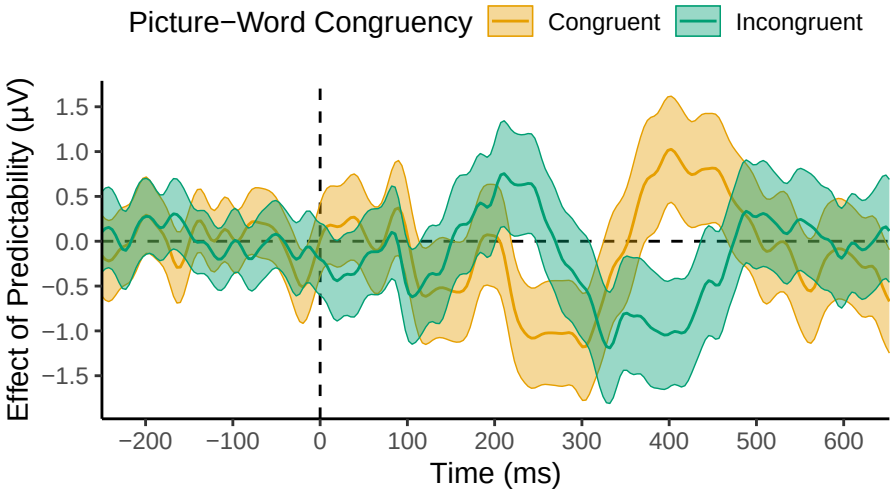


(**a**) Time-course of fixed effects estimates, with blue-shaded regions depicting 95% confidence intervals. The model intercept (reflecting amplitudes at the lowest level of Predictability) is depicted as a grey line on each panel to provide a reference for timing and magnitude of effects. (**b**) Fixed-effect predictions for picture-congruent and -incongruent words at levels of Predictability from 10 to 100%, in steps of 10%. (**c**) Same data as (**b**), but split by Predictability rather than Congruency.

reduced amplitudes in the N1 for picture-incongruent words, but increased amplitudes for picture-congruent words. This difference peaked around 225 ms, but reversed in direction after 300 ms. It is of note that the peak of the observed effects in the N1 was later than originally anticipated (the planned analysis was limited to ≤200 ms). Nevertheless, the model intercept (**Figure 12a**) clearly shows that these effects peaked during the N1's offset period.

Figure 13

Time-course of the effect of Predictability for picture-congruent and -incongruent words.



Central lines depict effect estimates, derived from sample-level models that were coded such that the model intercept lay at the respective levels of picture-word Congruency. Estimates reflect ERPs for words at the maximum level of Predictability, minus those at the minimum level of Predictability. Shaded areas depict 95% confidence intervals of model estimates.

Discussion

In the present study, we tested whether a simple predictive coding account could explain online prediction effects on the amplitude of N1 ERP components elicited by words in biasing contexts. We biased expectations for upcoming words via images of varying predictability. Based on a predictive coding framework, we hypothesised that there would be an interaction between picture-word Predictability and Congruency in which N1 amplitude scales with prediction error. Planned analyses failed to find evidence for this hypothesis, and exploratory analyses revealed, despite strong evidence for prediction effects

in the N1, that the direction of the interaction was opposite to that expected under the hypothesis. Specifically, increases in Predictability were associated with greater-amplitude N1s for picture-congruent words, and smaller-amplitude N1s for picture-incongruent words. On this basis, we conclude that a simple predictive coding explanation of the N1 cannot explain predictability effects observed in the picture-word verification task used here.

In recent years, predictive coding models have been increasingly applied to explain neural phenomena observed during language processing. This includes predictive coding perspectives on the N1 specifically (e.g., Gagl et al., 2020; Huang et al., 2022; Zhao et al., 2019), or its likely generator, vOT (Price & Devlin, 2011), and other areas of language processing. For example, consider the well-researched N400 ERP component, generally recognised since its initial identification as capturing activity related to semantic processes (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). The N400 shows sensitivity to word- and sentence-level surprise or predictability (Delaney-Busch et al., 2019; Lau et al., 2013; Lindborg et al., 2023; Mantegna et al., 2019; Van Petten & Kutas, 1990), in a manner that may be consistent with predictive coding (Bornkessel-Schlesewsky & Schlesewsky, 2019; Eddine et al., 2023; Rabovsky & McRae, 2014). Similar interpretations have been made of other signals, as capturing prediction errors for phonological, semantic, or syntactic representations (Fitz & Chang, 2019; Gagnepain et al., 2012; Van Petten & Luka, 2012; Ylinen et al., 2017; Ylinen et al., 2016). Indeed, emerging evidence supports the broader contention that naturalistic language comprehension utilises a predictive coding hierarchy spanning the language network (Caucheteux et al., 2023; Schuster et al., 2021; Shain et al., 2020).

We do not believe our findings refute the existence of predictive coding mechanisms during the N1. Instead, we argue that a simple predictive coding account of the N1, in which the component's amplitude straightforwardly indexes prediction error in a manner dependent on prediction certainty, is insufficient to explain the pattern of effects we observed in the picture-word verification task we used here. For a predictive coding model

to better account for these data, it would require elaboration. One feature that may be relevant is the nature of the task. We elected to use a picture-word verification task as it encourages explicit prediction of word forms from non-linguistic contexts. However, this task paradigm may alter predictive processing of word forms in two key ways. First, participants will have soon learned that the observed word form only matches its preceding image 50% of the time, which could have interacted with the effect of Predictability (prediction certainty) in unexpected ways. Second, the requirement for explicit verification of prediction congruency may have encouraged artificial processing strategies that are not representative of naturalistic word recognition and reading processes. To better understand whether and how such factors influence any possible predictive coding effects on the N1, we could manipulate prediction error magnitude and precision while the participant's task instructions do not explicitly require processing of the cue. For instance, we could use a picture-word priming design (Sperber et al., 1979; Vanderwart, 1984), presenting picture-word pairs, as in the current study, but ask participants to respond with lexical decisions. Here, prediction error magnitude could be operationalised as the orthographic distance between the string (whether word or non-word), and precision as the predictability of a word given its picture. We believe that such an approach could provide insight into whether, and which, features of the paradigm we used could have resulted in the unexpected pattern of results. Finally, it is possible that dynamics of predictive processing were influenced by the slow presentation rate employed in the present study, relative to more naturalistic reading paradigms. Indeed, previous research has highlighted the importance of presentation rate in prediction effects during reading (e.g., Dambacher et al., 2012), and recent findings have shown that unpredictability in stimulus presentation timing (e.g., with jittered inter-stimulus intervals) may interfere with predictive processes, as indexed by the mismatch negativity component (Tsogli et al., 2022). This explanation of our results could be tested by study designs examining how the congruency-predictability interaction varies over stimulus onset asynchronies of different durations. In sum, while

predictive coding mechanisms may ultimately underlie the pattern of effects we observed, the simple account we have tested requires elaboration, informed by insights from other paradigms, for it to explain why our current pattern of effects is opposite to that expected.

Nevertheless, we acknowledge the possibility that the insufficiency of predictive coding accounts to explain the data we observed may reflect a more fundamental shortcoming. To speculate, predictive coding models may account for activity in the N1 in previously tested paradigms without accurately describing the underlying neural processes. For instance, Luthra et al. (2021) showed that, in spoken word recognition, interactive activation models may provide an alternative account of the ERP amplitude reduction observed in response to prediction violations, without invoking key features of predictive coding models. Indeed, effects indicative of predictive processing may emerge in a system that that lacks any representations of, or mechanisms implementing, predictions or prediction errors, instead only implementing pattern completion (Falandays et al., 2021). It is tentatively possible that the picture-word verification paradigm we applied here may be a scenario that employs the same neurocognitive processes in the N1 as those employed in other paradigms, but elicits cognitive dynamics whose corresponding neural activity reveals differences from a predictive coding model. It is possible that processing indexed by the N1 can only be explained by a model distinct from the predictive coding framework, even though predictive coding models may correlate with patterns of activity seen in most paradigms. Justifying the development of such a model, distinct from predictive coding, would require much more evidence for the shortcomings of a predictive coding account, and we do not believe our study provides the insights necessary to speculate on the form such a model could take.

If a predictive coding account is to explain prediction-driven modulation of activity in the N1, or any component, we believe it is vital for researchers to consider the informational content of representations whose processing is indexed by the component which is thought to capture prediction error. In a hierarchical model of predictive coding,

where levels of the hierarchy utilise different representational formats, the interaction between ascending input and descending predictions must involve some mapping of higher-level onto lower-level representations. For instance, if semantic context can influence processing that is closer to sensory input and indexed by early ERP components (e.g., Enge et al., 2023; Getz & Toscano, 2019; Segalowitz & Zheng, 2009), then higher-level semantic information must be translated into predictions of upcoming lower-level sensory signals. In the case of our study's modulation of the N1, where the N1 is largely implicated in visual-orthographic processing (Bentin et al., 1999; Brem et al., 2018; Ling et al., 2019; Maurer, Brandeis, et al., 2005), predictions of upcoming words must be translated into a visual-orthographic code. Such a mapping could be expected to be very computationally lossy; predictions for visual-orthographic features of a single word should be expected to also confer facilitation for words that are orthographically similar, yet picture-incongruent.

From one perspective, mapping of predictions to lower-level representations may be considered a requisite for a phenomenon to be considered top-down modulation (Rauss et al., 2011). This relates to a long-standing debate on whether prediction effects at the lexical level of language processing necessitate top-down input informed by higher-level semantic processes, or could instead result from perhaps more parsimonious intralexical effects (Fodor, 1983; Forster, 1979). A similar argument could be made that context effects on the N1 could be interpreted as intra-*orthographic*, resulting from local interactions in a possible *orthographic module*. As an example, the orthographic features of the word form *fish* may preactivate features of the word form *chips* simply through learned co-occurrence rather than top-down modulation, entirely within an orthographic processing module that possesses nothing approaching a semantic representation. Such facilitation could be implemented via an extension to classic interactive activation models (e.g., McClelland & Rumelhart, 1981) in which there are excitatory lateral connections between word-level units whose strength is determined by co-occurrence frequency. We consider this point to highlight an advantage of paradigms such as ours, that use non-linguistic contexts (e.g.,

task instructions, images, etc.) to cue upcoming words and word forms. Effects of context that map across representations in this way necessitate transfer of information across levels of the processing hierarchy, and may thus be considered stronger evidence for an influence of top-down predictions.

An aspect of the predictive coding account that our design did not fully test also relates to this idea of representational mapping. We dichotomised the variable of congruency (prediction error magnitude), with orthographic Levenshtein distance maximised between picture-congruent and -incongruent word forms. However, prediction error magnitude should also be expected to vary continuously, from unpredicted word forms that are less to more orthographically similar to the predicted word form. This is comparable to Gagl et al.'s (2020) use of a pixel distance metric to calculate the continuous distance between a presented word form and a context-neutral prior. Such an approach could be applied to biasing contexts by instead calculating the orthographic distance between a presented word form and a context-informed prior, where the probability of observing certain pixels (or orthographic features) could be up-weighted proportional to prediction certainty. We believe such an approach could provide useful insights in elucidating the pattern of effects we observed.

In sum, we tested a simple predictive coding account of the word-elicited N1, but failed to find evidence in favour of it. Exploratory analyses suggest that the pattern of effects in the Congruency-Predictability interaction were in the opposite direction to that expected under a simple predictive coding model. We argue that such a model is insufficient to explain the pattern of effects we observed, and we have identified avenues of future research that could better delineate how predictive processes interact with processing during the N1.

References

Allison, T., Mccarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex*, *4*(5), 544–554. https://doi.org/10.1093/cercor/4.5.544

Appelbaum, L. G., Liotti, M., Perez, R., Fox, S. P., & Woldorff, M. G. (2009). The temporal dynamics of implicit processing of non-letter, letter, and word-forms in the human visual cortex. *Frontiers in Human Neuroscience*, *3*, 1–11. https://doi.org/10.3389/neuro.09.056.2009

Barr, D. J., Levy, J., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(Supplement), 1–63. https://doi.org/10.1016/j.jml.2012.11.001.Random

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260. https://doi.org/10.1162/089892999563373

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology*, *10*(Article 298). https://doi.org/10.3389/fpsyg.2019.00298

Brem, S., Halder, P., Bucher, K., Summers, P., Martin, E., & Brandeis, D. (2009). Tuning of the visual word processing system: Distinct developmental ERP and fMRI effects. *Human Brain Mapping*, *30*(6), 1833–1844. https://doi.org/10.1002/hbm.20751

Brem, S., Hunkeler, E., Mächler, M., Kronschnabel, J., Karipidis, I. I., Pleisch, G., & Brandeis, D. (2018). Increasing expertise to a novel script modulates the visual N1

ERP in healthy adults. *International Journal of Behavioral Development*, *42*(3), 333–341. https://doi.org/10.1177/0165025417727871

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS ONE*, *9*(9), 1–10. https://doi.org/10.1371/journal.pone.0106953

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 1–12. https://doi.org/10.1038/s41562-022-01516-2

Chang, C. Y., Hsu, S. H., Pion-Tonachini, L., & Jung, T. P. (2020). Evaluation of Artifact Subspace Reconstruction for automatic artifact components removal in multi-channel EEG recordings. *IEEE Transactions on Biomedical Engineering*, *67*(4), 1114–1121. https://doi.org/10.1109/TBME.2019.2930186

Chen, Y., Davis, M. H., Pulvermüller, F., & Hauk, O. (2013). Task modulation of brain responses in visual word recognition as studied using EEG/MEG and fMRI. *Frontiers in Human Neuroscience*, *7*, 1–14. https://doi.org/10.3389/fnhum.2013.00376

Chen, Y., Davis, M. H., Pulvermüller, F., & Hauk, O. (2015). Early visual word processing is flexible: Evidence from spatiotemporal brain dynamics. *Journal of Cognitive Neuroscience*, *27*(9), 1738–1751. https://doi.org/10.1162/jocn_a_00815

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*, 291–307.

Compton, P. E., Grossenbacher, P., Posner, M. I., & Tucker, D. M. (1991). A cognitive-anatomical approach to attention in lexical access. *Journal of Cognitive Neuroscience*, *3*(4), 304–312. https://doi.org/10.1162/jocn.1991.3.4.304

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Dambacher, M., Dimigen, O., Braun, M., Wille, K., Jacobs, A. M., & Kliegl, R. (2012). Stimulus onset asynchrony and the timeline of word recognition: Event-related potentials during sentence reading. *Neuropsychologia*, *50*(8), 1852–1870. https://doi.org/10.1016/j.neuropsychologia.2012.04.011

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). Measuring the associative structure of English: The "Small World of Words" norms for word association. *Behavior Resarch Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20. https://doi.org/10.1016/j.cognition.2019.01.001

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

Dikker, S., & Pylkkanen, L. (2011). Before the N400: Effects of lexical-semantic violations in visual cortex. *Brain and Language*, *118*(1-2), 23–28. https://doi.org/10.1016/j.bandl.2011.02.006

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2020). *GNU Octave version 6.1.0 manual: A high-level interactive language for numerical computations*. https://www.gnu.org/software/octave/doc/v6.1.0/

Eberhard-Moscicka, A. K., Jost, L. B., Fehlbaum, L. V., Pfenninger, S. E., & Maurer, U. (2016). Temporal dynamics of early visual word processing - Early versus late N1 sensitivity in children and adults. *Neuropsychologia*, *91*, 509–518. https://doi.org/10.1016/j.neuropsychologia.2016.09.014

Eddine, S. N., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2023). A predictive coding model of the N400. https://doi.org/10.1101/2023.04.10.536279

Enge, A., Süß, F., & Rahman, R. A. (2023). Instant effects of semantic information on visual perception. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2038-22.2023

Falandays, J. B., Nguyen, B., & Spivey, M. J. (2021). Is prediction nothing more than multi-scale pattern completion of the future? *Brain Research*, *1768*, 147578. https://doi.org/10.1016/j.brainres.2021.147578

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fodor, J. (1983). Input systems as modules. In *The Modularity of Mind* (pp. 47–101). MIT Press.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. Cooper & E. Walker (Eds.), *Sentence Processing: Psycholinguistic essays presented to Merrill Garrett* (pp. 27–85). Erlbaum.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gagl, B., Sassenhagen, J., Haan, S., Gregorova, K., Richlan, F., & Fiebach, C. J. (2020). An orthographic prediction error as the basis for efficient visual word recognition. *NeuroImage*, *214*(August 2019), 116727. https://doi.org/10.1016/j.neuroimage.2020.116727

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, *30*(6), 830–841. https://doi.org/10.1177/0956797619841813

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Hauk, O., Coutout, C., Holden, A., & Chen, Y. (2012). The time-course of single-word reading: Evidence from fast behavioral and brain responses. *NeuroImage*, *60*(2), 1462–1477. https://doi.org/10.1016/j.neuroimage.2012.01.061

Huang, Z., Yang, S., Xue, L., Yang, H., Lv, Y., & Zhao, J. (2022). Level of orthographic knowledge helps to reveal automatic predictions in visual word processing. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.809574

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492. https://doi.org/10.1162/neco.1997.9.7.1483

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668). https://doi.org/10.1098/rstb.2014.0169

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Kherif, F., Josse, G., & Price, C. J. (2011). Automatic top-down processing explains common left occipito-temporal responses to visual words and objects. *Cerebral Cortex*, *21*, 103–114. https://doi.org/10.1093/cercor/bhq063

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from ERPs. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112. https://doi.org/10.1162/jocn_a_00148

Kim, A. E., & Gilley, P. M. (2013). Neural mechanisms of rapid sensitivity to syntactic anomaly. *Frontiers in Psychology*, *4*, 1–15. https://doi.org/10.3389/fpsyg.2013.00045

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

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

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

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

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502. https://doi.org/10.1162/jocn_a_00328

Lin, S. E., Chen, H. C., Zhao, J., Li, S., He, S., & Weng, X. C. (2011). Left-lateralized N170 response to unpronounceable pseudo but not false Chinese characters - the key role of orthography. *Neuroscience*, *190*, 200–206. https://doi.org/10.1016/j.neuroscience.2011.05.071

Lindborg, A., Musiolek, L., Ostwald, D., & Rabovsky, M. (2023). Semantic surprise predicts the N400 brain potential. *Neuroimage: Reports*, *3*(1), 100161. https://doi.org/10.1016/j.ynirp.2023.100161

Ling, S., Lee, A. C. H., Armstrong, B. C., & Nestor, A. (2019). How are visual words represented? Insights from EEG-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, *40*(17), 5056–5068. https://doi.org/10.1002/hbm.24757

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Luthra, S., Li, M. Y. C., You, H., Brodbeck, C., & Magnuson, J. S. (2021). Does signal reduction imply predictive coding in models of spoken word recognition?

Psychonomic Bulletin & Review, *28*(4), 1381–1389. https://doi.org/10.3758/s13423-021-01924-x

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

MATLAB. (2022). *MATLAB Version 9.13.0 (R2022b)*. The MathWorks Inc.

Maurer, U., Brandeis, D., & McCandliss, B. D. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. *Behavioral and Brain Functions*, *1*, 1–12. https://doi.org/10.1186/1744-9081-1-13

Maurer, U., Brem, S., Bucher, K., & Brandeis, D. (2005). Emerging neurophysiological specialization for letter strings. *Journal of Cognitive Neuroscience*, *17*(10), 1532–1552. https://doi.org/10.1162/089892905774597218

Maurer, U., Rossion, B., & McCandliss, B. D. (2008). Category specificity in early perception: Face and word N170 responses differ in both lateralization and habituation properties. *Frontiers in Human Neuroscience*, *2*, 1–7. https://doi.org/10.3389/neuro.09.018.2008

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An acccount of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews*, *96*, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Nobre, A. C., Allison, T., & McCarthy, G. (1994). Word recognition in the human inferior temporal lobe. *372*, 260–263. https://doi.org/10.1038/372260a0

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, *10*(1089), 1–8. https://doi.org/10.3389/fpsyg.2019.01089

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, *74*(3), 374–388. https://doi.org/10.1016/j.biopsycho.2006.09.008

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Pleisch, G., Karipidis, I. I., Brem, A., Röthlisberger, M., Roth, A., Brandeis, D., Walitza, S., & Brem, S. (2019). Simultaneous EEG and fMRI reveals stronger sensitivity to orthographic strings in the left occipito-temporal cortex of typical versus poor beginning readers. *Developmental Cognitive Neuroscience*, *40*, 1–13. https://doi.org/10.1016/j.dcn.2019.100717

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26–46. https://doi.org/10.1.1.443.7693

Price, C. J., & Devlin, J. T. (2011). The Interactive Account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*(6), 246–253. https://doi.org/10.1016/j.tics.2011.04.001

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rauschecker, A. M., Bowen, R. F., Parvizi, J., & Wandell, B. A. (2012). Position sensitivity in the visual word form area. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24). https://doi.org/10.1073/pnas.1121304109

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

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 514–528. https://doi.org/10.1037/a0020990

Rodrigues, A. P., Rebola, J., Pereira, M., Van Asselen, M., & Castelo-Branco, M. (2019). Neural responses of the anterior ventral occipitotemporal cortex in developmental dyslexia: Beyond the visual word form area. *Investigative Ophthalmology and Visual Science*, *60*(4), 1063–1068. https://doi.org/10.1167/iovs.18-26325

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

Schuster, S., Himmelstoss, N. A., Hutzler, F., Richlan, F., Kronbichler, M., & Hawelka, S. (2021). Cloze enough? Hemodynamic effects of predictive processing during natural reading. *NeuroImage*, *228*, 117687. https://doi.org/10.1016/j.neuroimage.2020.117687

Segalowitz, S. J., & Zheng, X. (2009). An ERP study of category priming: Evidence of early lexical semantic access. *Biological Psychology*, *80*(1), 122–129. https://doi.org/10.1016/j.biopsycho.2008.04.009

Sereno, S. C., Brewer, C. C., & O'Donnell, P. J. (2003). Context effects in word recognition: Evidence for early interactive processing. *Psychological Science*, *14*(4), 328–333. https://doi.org/10.1111/1467-9280.14471

Sereno, S. C., Hand, C. J., Shahid, A., Mackenzie, I. G., & Leuthold, H. (2019). Early EEG correlates of word frequency and contextual predictability in reading [Publisher: Taylor & Francis]. *Language, Cognition and Neuroscience*, *35*(5), 625–640. https://doi.org/10.1080/23273798.2019.1580753

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport*, *9*(10), 2195–2200. https://doi.org/10.1097/00001756-199807130-00009

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition*, *7*(5), 339–345. https://doi.org/10.3758/BF03196937

STAN Development Team. (2023). Stan Modeling Language Users Guide and Reference Manual, 2.32. https://mc-stan.org

Strijkers, K., Bertrand, D., & Grainger, J. (2015). Seeing the same words differently: The time course of automaticity and top-down intention in reading. *Journal of Cognitive Neuroscience*, *27*(8), 1542–1551. https://doi.org/10.1162/jocn_a_00797

Taha, H., Ibrahim, R., & Khateb, A. (2013). How does arabic orthographic connectivity modulate brain activity during visual word recognition: An ERP study. *Brain Topography*, *26*(2), 292–302. https://doi.org/10.1007/s10548-012-0241-2

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

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*, 2372–2382. https://doi.org/10.3758/s13428-020-01389-1

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Tsogli, V., Jentschke, S., & Koelsch, S. (2022). Unpredictability of the "when" influences prediction error processing of the "what" and "where" [Publisher: Public Library of Science]. *PLOS ONE*, *17*(2), e0263373. https://doi.org/10.1371/journal.pone.0263373

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Vanderwart, M. (1984). Priming by pictures in lexical decision. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 67–83. https://doi.org/10.1016/S0022-5371(84)90509-7

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

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

Veale, J. F. (2014). Edinburgh Handedness Inventory - Short Form: A revised version based on confirmatory factor analysis. *Laterality*, *19*(2), 164–177. https://doi.org/10.1080/1357650X.2013.783045

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels Artificial Character Sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods*, *49*(6), 2093–2112. https://doi.org/10.3758/s13428-016-0844-8

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Wang, F., & Maurer, U. (2017). Top-down modulation of early print-tuned neural activity in reading. *Neuropsychologia*, *102*, 29–38. https://doi.org/10.1016/j.neuropsychologia.2017.05.028

Wang, F., & Maurer, U. (2020). Interaction of top-down category-level expectation and bottom-up sensory input in early stages of visual-orthographic processing. *Neuropsychologia*, *137*, 107299. https://doi.org/10.1016/j.neuropsychologia.2019.107299

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

White, A. L., Palmer, J., Boynton, G. M., & Yeatman, J. D. (2019). Parallel spatial channels converge at a bottleneck in anterior word-selective cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 10087–10096. https://doi.org/10.1073/pnas.1822137116

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398. https://doi.org/10.1038/s41562-020-00982-w

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yeatman, J. D., Rauschecker, A. M., & Wandell, B. A. (2013). Anatomy of the visual word form area: Adjacent cortical circuits and long-range white matter connections. *Brain and Language*, *125*(2), 146–155. https://doi.org/10.1016/j.bandl.2012.04.010

Ylinen, S., Bosseler, A., Junttila, K., & Huotilainen, M. (2017). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science*, *20*(6), e12472. https://doi.org/10.1111/desc.12472

Ylinen, S., Huuskonen, M., Mikkola, K., Saure, E., Sinkkonen, T., & Paavilainen, P. (2016). Predictive coding of phonological rules in auditory cortex: A mismatch negativity study. *Brain and Language*, *162*, 72–80. https://doi.org/10.1016/j.bandl.2016.08.007

Zhao, J., Kipp, K., Gaspar, C., Maurer, U., Weng, X., Mecklinger, A., & Li, S. (2014). Fine neural tuning for orthographic properties of words emerges early in children reading

alphabetic script. *Journal of Cognitive Neuroscience*, *26*(11), 2431–2442. https://doi.org/10.1162/jocn_a_00660

Zhao, J., Maurer, U., He, S., & Weng, X. (2019). Development of neural specialization for print: Evidence for predictive coding in visual word recognition. *PLoS Biology*, *17*(10), 1–17. https://doi.org/10.1371/journal.pbio.3000474