AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General


ISSN: 0096-3445

2025, Vol. 154, No. 3, 711–738
https://doi.org/10.1037/xge0001677

# Let Them Eat *Ceke*: An Electrophysiological Study of Form-Based Prediction in Rich Naturalistic Contexts

Anthony Yacovone[1, 2], Briony Waite[1], Tatyana Levari[1], and Jesse Snedeker[1]
[1] Department of Psychology, Harvard University
[2] Department of Linguistics, Boston University

It is well-established that people make predictions during language comprehension—the nature and specificity of these predictions, however, remain unclear. For example, do comprehenders routinely make predictions about which words (and phonological forms) might come next in a conversation, or do they simply make broad predictions about the gist of the unfolding context? Prior EEG studies using tightly controlled experimental designs have shown that form-based prediction can occur during comprehension, as N400s to unexpected words are reduced when they resemble the form of a predicted word (e.g., *ceke* when expecting *cake*). One limitation, however, is that these studies often create environments that are optimal for eliciting form-based prediction (e.g., highly constraining sentences, slower-than-natural rates of presentation). Thus, questions remain about whether form-based prediction can occur in settings that more closely resemble everyday comprehension. To address this, the present study explores form-based prediction during naturalistic spoken language comprehension. English-speaking adults listened to a story in which some of the words had been altered. Specifically, we experimentally manipulated whether participants heard the original word from the story (*cake*), a form-similar nonword (*ceke*), or a less-similar nonword (*vake*). Half of the target words were predictable given their context, and the other half were unpredictable. Consistent with the prior work, we found reduced N400s for form-similar nonwords (*ceke*) relative to less-similar nonwords (*vake*)—but only in predictable contexts. This study demonstrates that form-based prediction can emerge in naturalistic contexts, and therefore, it is likely to be a common aspect of language comprehension in the wild.

***Public Significance Statement***
This study demonstrates that English-speaking adults predict the initial sounds of upcoming words when listening to stories. Form-based prediction of this kind has long been considered a marginal phenomenon in language comprehension research. Thus, these findings suggest that, by using ecologically valid techniques like the *Storytime* paradigm, researchers can better understand the mechanisms that humans use to process information in their everyday lives.

*Keywords:* language comprehension, form-based prediction, naturalistic discourse, event-related potentials, N400



This article was published Online First December 16, 2024.
Nicole Wicha served as action editor.
Anthony Yacovone https://orcid.org/0000-0001-5151-4472
Briony Waite https://orcid.org/0009-0001-2730-2522
Tatyana Levari https://orcid.org/0000-0002-3523-3117
Jesse Snedeker https://orcid.org/0000-0002-3658-2888

All preregistrations, data, code, and study materials are publicly available on the Open Science Framework (OSF; https://osf.io/upjc4/). Aspects of this research have been presented at professional conferences and discussed in Anthony Yacovone's doctoral dissertation. The authors have no conflicts of interest to disclose.

This work was funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health (1R03HD097629-01 awarded to Jesse Snedeker), and Harvard Restricted Funds (Anthony Yacovone). The authors thank many research assistants who helped with data collection, especially Paulina Piwowarczyk and Moshe Poliak. The authors also thank Kathryn Davidson, Roger Levy, and Gina Kuperberg for their feedback on earlier drafts and their technical guidance on data analysis and interpretation.

Anthony Yacovone played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, and writing–review and editing, and a supporting role in funding acquisition. Briony Waite played a supporting role in data curation, investigation, methodology, project administration, resources, and writing–review and editing. Tatyana Levari played a supporting role in conceptualization, funding acquisition, software, and writing–review and editing. Jesse Snedeker played a lead role in funding acquisition, a supporting role in formal analysis, project administration, and resources, and an equal role in conceptualization, methodology, supervision, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Anthony Yacovone, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, United States. Email: anthony_yacovone@g.harvard.edu

When listening to a story or conversation, we use the sounds that we are hearing to reconstruct the message that the speaker is trying to convey. To do this, we must represent the incoming signal at multiple distinct levels (as phonemes, words, syntactic structures, and ideas). One of the central discoveries in psycholinguistics is that these representations are built (and refined) via both bottom-up and top-down processing. Bottom-up processing is when information from one level is used to build higher, more abstract representations (e.g., turning sounds into words and words into phrases). In contrast, top-down processing is when information from higher levels is used to influence which representations are being built at levels below (e.g., using world knowledge to interpret what someone just said or even to predict what someone is about to say).

Decades of research in psycholinguistics have focused on the role of top-down processing during language comprehension—and in particular, the role of linguistic prediction.[1] This work provides ample evidence for predictive processing of this kind, largely from reading time studies, visual world eye-tracking studies, and studies that use electroencephalography (EEG). In these studies, comprehenders read predictable words faster than unpredictable ones; they look toward particular referents in anticipation of them being mentioned; and they show reduced neural responses to words that are consistent with their top-down predictions (for reading studies, see Ehrlich & Rayner, 1981; Rayner & Well, 1996; Smith & Levy, 2013; for visual world studies, see Altmann & Kamide, 1999; Borovsky et al., 2012; Kamide et al., 2003; Milburn et al., 2016; and for EEG studies, see DeLong et al., 2005; Federmeier & Kutas, 1999; Van Berkum et al., 2005; Wicha et al., 2004).

Given these findings, there is now a general consensus that top-down prediction occurs during language comprehension (Pickering & Gambi, 2018). What remains unclear is whether prediction occurs at *all* levels of representation or only at higher ones (DeLong et al., 2021; Freunberger & Roehm, 2016; Ito et al., 2016; Nieuwland, 2019). One possibility is that, in general, comprehenders only make broad, high-level predictions about the gist of a sentence's meaning as it unfolds. The other possibility is that, in addition to making these more general predictions, comprehenders might also make more precise predictions about the specific word(s) that will come next (Altmann & Mirković, 2009; Heilbron et al., 2022; Willems et al., 2016). Recent evidence from EEG suggests that comprehenders' predictions can be lexically specific and occur at both levels of word meaning and word form (Brothers et al., 2015; DeLong et al., 2005; Ito et al., 2016; Laszlo & Federmeier, 2009; Wicha et al., 2004). But as we discuss below, form-based prediction appears to be more limited than semantic prediction, occurring primarily in tightly controlled experiments that use very predictable designs and/or slower-than-natural rates of presentation (for discussion, see Ito et al., 2016; but see DeLong et al., 2021). In the present study, we try to better understand the scope of this phenomenon by asking whether form-based prediction occurs when people simply listen to a naturally produced story with no explicit task beyond understanding it.

In the remainder of this introduction, we will do three things: First, we review the EEG literature on form-based prediction during language comprehension. Second, we consider the paradigms that are typically used, the limits of their ecological validity, and the questions that these limits raise about prediction in the wild. Finally, we discuss how the present study is designed to explore form-based prediction in a more naturalistic listening context.

## Reviewing the EEG Evidence for Form-Based Prediction During Comprehension

EEG studies have been central to our understanding of predictive processes during language comprehension (e.g., Beres, 2017; Federmeier, 2022; Kutas et al., 2006, 2014; Kutas & Federmeier, 2011; Payne et al., 2020; Swaab et al., 2012; Van Petten & Luka, 2012). In these studies, researchers typically record changes in participants' neural activity at the scalp as they comprehend a variety of sentences (Kutas et al., 2006; Kutas & Van Petten, 1994; Morgan-Short & Tanner, 2013; Swaab et al., 2012). These recordings are then time-locked to the onset of particular words, creating event-related potentials (ERPs). The interpretation of ERPs focuses on stable patterns of neural activity called *components*, which are typically distinguished from one another based on their voltage direction, peak latency, and scalp distribution, as well as their sensitivity to particular variables (for reviews, see Kappenman & Luck, 2011; Luck, 2014). In the following sections, we review the ERP components that most commonly emerge in the study of form-based prediction in spoken language: the N400, the P600, and two early negativities known as the Phonological Mismatch Negativity (PMN) and the N200.

### *The N400 as an Effect of Lexicosemantic Preactivation During Comprehension*

One of the best understood and most replicated components in psycholinguistic research is the N400. This component is a negative-going deflection in the ERP waveform that typically emerges over centroparietal electrode sites and peaks between 300 and 500 ms poststimulus onset (Kutas & Federmeier, 2011). The N400 was first observed in studies that had participants read sentences with anomalous endings, for example, "He spread the warm bread with *socks*" (Kutas & Hillyard, 1980). For this reason, the N400 was initially characterized as a response to semantic anomalies. However, most contemporary theorists reject this characterization because it fails to account for the wide range of *plausible* contexts in which N400 effects also appear (Federmeier, 2007, 2022; Federmeier et al., 2007; Kuperberg, 2007; Kuperberg et al., 2020; Kutas & Federmeier, 2011). In fact, there is an N400 response to *every* word, whether it is presented in isolation or within a sentence context (Kutas, 1993; Kutas & Federmeier, 2000, 2011; Payne et al., 2015; Rugg, 1990; Van Petten & Kutas, 1990).

In contemporary psycholinguistic theories, the N400 is seen as an index of the relative ease of accessing the lexicosemantic features of a word (e.g., Federmeier, 2022; Kuperberg et al., 2020). There is ample evidence to support this interpretation: First, the N400 for

[1] In the present study, we consider any preactivation due to top-down processing as a form of prediction (for similar definitions, see DeLong, Troyer, et al., 2014; DeLong et al., 2021; Huettig et al., 2022; Kuperberg & Jaeger, 2016, Section 3, p. 39; Kutas & Federmeier, 2011; Pickering & Gambi, 2018). This contrasts with theorists who reserve the term *prediction* for a distinct form of processing in which an active commitment is made to upcoming material using mechanisms that are distinct from the incremental, top-down processes described above (see Kuperberg & Jaeger, 2016, Section 4, pp. 45–47; Kutas et al., 2011, for discussion).

a given word is larger when the word is presented in isolation and smaller when it is presented within a plausible sentence (Kutas, 1993). As we described above, when words are presented in a broader context, they often become predictable to some degree, and comprehenders may be able to preactivate representations associated with them before they appear in the input. Thus, the reduction in the N400 response to a word within a sentence is often attributed to top-down prediction facilitating lexicosemantic processing (Federmeier, 2007; Lau et al., 2008, 2013; for computational descriptions of the N400, see Nour Eddine et al., 2022). Second, within a given sentence, the N400 to each subsequent word decreases as the cumulative context makes each word more and more predictable (Payne et al., 2015; Van Petten & Kutas, 1990, 1991). Third, N400 responses have an inverse correlation with cloze probability measures from offline sentence completion tasks, such that the N400 responses to words become smaller as the predictability of the words increases (Kutas et al., 2019; Kutas & Hillyard, 1984).

Subsequent research has built on this basic insight, using the N400 to explore which features of a word (or concept) are preactivated by the context during language comprehension (e.g., DeLong et al., 2005, 2019, 2021; Federmeier & Kutas, 1999; Heilbron et al., 2022; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Otten & Van Berkum, 2008; Wang et al., 2020; Wicha et al., 2004). For example, in a foundational study by Federmeier and Kutas (1999), participants read sentences like "The yard was completely covered with a thick layer of dead leaves. Erica decided it was time to get out the (*rake*/*shovel*/*hammer*)." In this context, the word *rake* is highly predictable, presumably leading to the preactivation of its semantic features. By hypothesis, this should result in reduced N400s to *shovel* relative to *hammer* because the former shares more semantic features with *rake* (e.g., both are tools for yard work). As expected, the authors observed graded N400 responses such that expected words (*rake*) produced the smallest N400s, followed by semantically similar words (*shovel*), and then dissimilar words (*hammer*).

Evidence for preactivation of form features comes primarily from reading studies with violations that are orthographic neighbors of the predicted word. For example, Laszlo and Federmeier (2009) had participants read sentences with highly predictable endings like (1) and (2) below. In this study, the authors replaced these predictable sentence-final words with violations that either resembled or did not resemble the orthographic form of the original word. These orthographic violations were manipulated between items such that some sentences ended with form-similar conditions (*neighbors*, as in 1) and other sentences ended with dissimilar conditions (*nonneighbors*, as in 2). The authors also manipulated the lexical status of each violation, such that the violations were either unexpected words (*bark, clam*), nonwords (*pank*, *horm*), or illegal strings (*bxnk*, *rqck*).

1. *Neighbors*: Before lunch, he had to deposit his paycheck at the (*bank*/*bark*/*pank*/*bxnk*).

2. *Nonneighbors*: The genie was ready to grant his third and final (*wish*/*clam*/*horm*/*rqck*).

Similar to Federmeier and Kutas (1999), the authors observed graded N400 responses such that orthographic neighbors (regardless of lexical status) produced smaller N400s than nonneighbors. They interpreted this finding as evidence that, in sentences with strong contextual constraints, readers can rapidly predict upcoming words and preactivate their orthographic features. These form-based predictions then facilitate the processing of expected words and form-similar violations. These findings have been replicated in a handful of other studies using both real-word and nonword manipulations (DeLong et al., 2019, 2021; Kim & Lai, 2012; Liu et al., 2006). And critically, these findings have been linked to top-down prediction, as the reduction in N400s to form-similar words disappears when the original target word is less predictable (Ito et al., 2016).

### *The Posterior P600 as an Effect of Reprocessing Strong Violations of Expectation*

Studies on form-based prediction often report another ERP component known as the P600 (see DeLong et al., 2019, 2021; Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009).[2] This component typically emerges between 600 and 1,000 ms over posterior electrode sites in response to anomalies or strong violations of expectation (see DeLong, Quante, et al., 2014; Kuperberg, 2007; Kuperberg et al., 2020; Van De Meerendonk et al., 2009; Van Petten & Luka, 2012). In the study above, Laszlo and Federmeier (2009) found P600s in response to their orthographic violations. Moreover, these P600s were sensitive to their manipulations of lexical status and orthographic similarity. First, they found that illegal strings (*bxnk*, *rqck*) produced the largest P600s, followed by nonwords (*pank*, *horm*), and then unexpected words (*bark*, *clam*). Second, regardless of lexical status, the violations that closely resembled the form of the original target words produced larger P600s than the dissimilar violations (see also Kim & Lai, 2012).

The precise interpretation of the P600 is still debated because it is observed in a wide range of psycholinguistic studies using various syntactic, semantic, and phonological violations (see Kuperberg et al., 2020; Ryskin et al., 2021; Van Petten & Luka, 2012). Researchers seem to agree, however, that the P600 reflects the initial failure to incorporate the bottom-up input into one's higher level interpretation of the context, as well as the set of processes related to reprocessing that anomalous input (Brothers et al., 2020, 2022; Hagoort & Brown, 1999; Hahne & Friederici, 1999; Ito et al., 2016; Kim & Lai, 2012; Kuperberg et al., 2020; Laszlo & Federmeier, 2009; Osterhout et al., 1994, 2002; Osterhout & Holcomb, 1992; van de Meerendonk et al., 2010; Van De Meerendonk et al., 2009; Vissers et al., 2006). In line with this account, P600s tend to be larger in highly predictable contexts (Gunter et al., 2000; Ito et al., 2016; van de Meerendonk et al., 2010; Vissers et al., 2006) and in situations that promote deep comprehension (e.g., reading a discourse or listening to a narrative; see Brothers et al., 2020, 2022; Kuperberg et al., 2020).

According to this broad interpretation, the P600s in form-based prediction research could index comprehenders' attempts to gather more information about the nature of the violations, i.e., reflecting on whether they misperceived the input or whether someone produced a typo or speech error (Brothers et al., 2020, 2022; Kuperberg, 2007; Kuperberg et al., 2020; Van De Meerendonk et al., 2009; van Herten et al., 2005; Vissers et al., 2006). We return to these findings and interpretations in the General Discussion.

[2] This component has also been labeled as a Late Positive Component and posterior post-N400 positivity. For ease of discussion, we will use the term P600 in this article to refer to all of these findings.

### *Early Negativities as Evidence for Form-Based Prediction in Spoken Language Contexts*

Finally, form-based prediction research has also uncovered various ERP components that emerge *before* the N400 and P600 responses. These early components systematically differ across modalities, as some are only evoked by written language while others are only evoked by spoken language. Because the present study uses naturalistic speech, we will describe two of these early components from prior studies on spoken language comprehension: the Phonological Mismatch Negativity (PMN) and the N200. These early negativities, however, are difficult to replicate (Lewendon et al., 2020; Nieuwland, 2019; Poulton & Nieuwland, 2022), and many researchers simply interpret them as early emerging N400 effects (e.g., Van Petten et al., 1999). But for the sake of completeness, we review the evidence for these two components below.

The first early component is the PMN, which was first reported in Connolly and Phillips (1994). In this study, the authors explored whether the initial stages of phonological processing could be influenced by listeners' top-down expectations about upcoming words. To do this, they had English-speaking adults listen to various sentences that strongly constrained for a particular sentence-final word, e.g., "At night, the old woman locked the *door*." For some sentences, the authors kept the expected sentence-final words (*door*, 3a). For other sentences, they replaced the expected words with violations that overlapped with them in their phonological onsets (*eyes* → *icicles*, 3b), semantic features (*sink* → *kitchen*, 3c), or neither (*milk* → *nose*, 3d). Note, in the condition with different phonological onsets but similar semantic features (3c), the authors ensured that the violating word (*kitchen*) was always less predictable than the original sentence-final word (*sink*) from the target sentence.

3a. At night, the old woman locked the *door*. (Onset match, semantic match; *door*)

3b. Phil put some drops in his *icicles*. (Onset match, semantic mismatch; *eyes*)

3c. They left the dirty dishes in the *kitchen*. (Onset mismatch, semantic match; *sink*)

3d. Joan fed her baby some warm *nose*. (Onset mismatch, semantic mismatch; *milk*)

Based on their prior work, the authors expected to find two distinct negativities: one related to processing unexpected phonological features at around 200–300 ms (the PMN) and one related to processing unexpected semantic features (the N400). Results indicated early negativities for conditions with an unexpected phonological onset (3c and 3d) relative to those conditions with the expected onset (3a and 3b). Then, in a later time window, there were greater negativities for the two semantic mismatch conditions (3b and 3d) relative to the conditions with semantically congruent endings (3a and 3c). Taken at face value, this pattern suggests that there are two categorically distinct effects: an early PMN and a later N400.

But other features of the data pattern suggest that the effects are not discrete. The early negativity in the double mismatch condition (*milk* → *nose*, 3d) was greater than the negativity from the condition with just a different onset (*sink* → *kitchen*, 3c), suggesting that features beyond phonology influenced this early negativity. In fact, in the double mismatch condition, the PMN and the N400 blurred together to form one large, broadly distributed negativity that lasted from roughly 200 to 600 ms.

These findings are often contrasted with another set of findings from Van Petten et al. (1999). In this study, English-speaking participants listened to constraining sentences like "It was a pleasant surprise to find that the car repair bill was only seventeen *dollars*." The authors manipulated the sentence-final word to be either the expected word (*dollars*), a semantically incongruous word that shares initial phonemes with the expected word (*dolphins*), or a semantically incongruous word that rhymes with the expected word (*scholars*). Results showed increased negativities for both incongruous conditions, which the authors labeled as N400 effects; however, similar to the findings above, the negativity for the condition with an unexpected onset (*scholars*) emerged earlier than the negativity for the condition with an expected onset (*dolphins*). Also, as in Connolly and Phillips (1994), there was only one broadly distributed negativity (rather than two distinct effects) for the condition with the unexpected onset and unexpected meaning (*scholars*).

The second early component is the N200, which is best characterized by van den Brink et al. (2001). In this study, Dutch-speaking adults listened to sentences with highly predictable sentence-final words, e.g., "De schilder kleurde de details in met een klein *penseel*" (English translation, "The painter colored the details with a small paint *brush*"). Similar to Van Petten et al. (1999), the authors manipulated the target words to be either the expected word (*penseel*, "brush"), a semantically anomalous word with the same initial phonemes (*pensioen*, "pension"), or a semantically anomalous word with different initial phonemes (*doolhof*, "labyrinth"). In contrast with Van Petten et al. (1999), the authors found two distinct negative peaks in the ERP waveforms for all three conditions—one around 200 ms (the N200) and the other around 400 ms (the N400). The double mismatch condition (*doolhof*) produced a larger N200 than the two conditions with the expected initial phonemes (*penseel*, *pensioen*), which both produced similarly small N200 responses. The N400 effects showed a pattern similar to those found in the written studies above: The double mismatch violation (i.e., the phonologically dissimilar word, *doolhof*) had the largest N400; the shared onset violation (i.e., the phonologically similar word, *pensioen*) had a reduced N400 relative to the double mismatch violation; and finally, the expected word (*penseel*) had the smallest N400. These effects are slightly earlier than prior findings—but the authors argued that, because all of their target words began with plosives, their phonological effects may have been better aligned in time, producing more robust peaks in an earlier time window than seen in prior studies.

More recently, Boudewyn et al. (2015) reported an N200 effect in a study using story-based stimuli. Specifically, they had English-speaking adults listen to two-sentence discourses that were strongly constrained for a critical noun like *cake*. Then, they either presented the expected word (*cake*, average cloze probability = 78%) or an unexpected foil (*veggies*, average cloze probability = 0%). They also manipulated whether the immediate local context reinforced the global prediction (e.g., *sweet* and tasty… cake) or generated competing predictions (e.g., *healthy* and tasty … cake). As expected, there were graded N400s (300–600 ms) to the critical nouns such that the globally predicted nouns (regardless of local coherence) evoked the smallest N400s (sweet/healthy and tasty *cake*), followed by the globally unpredicted but locally coherent nouns (healthy and tasty *veggies*), and then the globally and locally

incoherent nouns (sweet and tasty *veggies*). In an early time window between 200 and 300 ms, Boudewyn et al. (2015) found increased frontal negativities for the globally and locally incoherent word (sweet and tasty *veggies*) relative to the other conditions. The authors interpreted this early N200 effect for the double violation condition as reflecting the *distinct* cost associated with detecting the mismatch between the predicted lexical form and the form that comprehenders encountered in the input.

Although some authors consider early negativities to be categorically distinct from the N400, others interpret such effects as early modulations of the N400 (e.g., Van Petten et al., 1999; for discussion, see Lewendon et al., 2020; Nieuwland, 2019). On these accounts, the negativities produced by the unexpected conditions in the studies above simply reflect the degree of facilitated access at the form level due to the prior preactivation of phonological features. There are two patterns that favor this hypothesis: First, as the findings above suggest, these early negativities vary considerably in their timing and scalp locations, and they are often continuous (in time and space) with the later N400 effects. Second, the same factors that influence the N400 also influence these early negativities, suggesting a similar functional interpretation for these effects (see Lewendon et al., 2020; Nieuwland, 2019). Given the wide range of interpretations, and the instability in these early ERP components, we focused our analyses on the N400 and P600 in the present study.

## Reviewing the Typical Paradigms in Form-Based Prediction Studies and Their Limitations

The findings above demonstrate that comprehenders can, under some circumstances, predict the form of upcoming words. Most of this evidence, however, comes from studies with paradigms that are very similar to one another and quite different from the typical contexts of language comprehension. Thus, it is unclear how broadly these findings generalize to more real-world settings. Specifically, all of the studies to date ask participants to attend to a stream of unrelated sentences with either no clear purpose or with a goal that is independent of comprehension (e.g., monitoring for errors). Often, but not always, these studies present their sentences in ways that diverge from how language is normally produced. For example, in many EEG reading studies, the sentences are presented word-by-word and often at a rate that is slower than typical reading (Ito et al., 2016; Kim & Lai, 2012; Laszlo & Federmeier, 2009; Vissers et al., 2006; but see DeLong et al., 2021).

Several recent studies have explored how presentation rate in particular might influence form-based prediction. For example, Ito et al. (2016) conducted two experiments in which they directly manipulated the predictability of sentences and the rates of presentation. Participants read sentences with highly predictable target words (e.g., "The student is going to the library to borrow a *book* tomorrow") and moderately predictable target words (e.g., "The family went to the sea to catch some *fish* together"). In these experiments, the target words were manipulated to be one of the following word types: the expected word (*book*), a semantically similar word with no form overlap (*page*), a semantically dissimilar word with form overlap (*hook*), or a word with no overlap in semantics or form (*sofa*). To test for an effect of presentation rate, they conducted the same experiment twice: In the first experiment, they presented sentences word-by-word with 500 ms between the onsets of each word—a rate similar to other studies in the literature (e.g., Kim & Lai, 2012; Laszlo & Federmeier, 2009). In the second experiment, the time between word onsets was increased to 700 ms.

As expected, Ito and colleagues found N400 effects in all violation conditions (*hook*, *page*, *sofa*) in both experiments. The N400 effects for semantically similar words (*page*) were smaller than those for unrelated words (*sofa*) at both presentation rates. This reduction, however, was only present in the highly predictable sentences, suggesting that preactivation of semantic features occurred regardless of presentation rate as long as the target word was predictable in its context. In contrast, the N400 effects for words with form overlap (*hook*) were only reduced when the sentence was both highly predictable *and* presented at a slower rate. In all other conditions, the N400 effects for words with overlapping forms (*hook*) patterned with those for the unrelated words (*sofa*). The authors concluded from these findings that it is easier to preactivate semantic features relative to lower level features (e.g., phonological form), perhaps because top-down predictions are initially made at the higher level of meaning and thus activate semantic representations before they can trickle down to lower levels and activate representations of form.

In short, Ito et al. (2016) found that, although semantic prediction occurred at both slower and faster presentation rates (1.5 words per second and two words per second, respectively), form-based prediction only occurred at the slower rate. Given that typical adults read three to five words per second (Brysbaert, 2019) and speak at a rate of about three to four words per second (Tauroza & Allison, 1990), many theorists have concluded that form-based prediction is unlikely to occur in most ordinary language comprehension contexts (Freunberger & Roehm, 2016, 2017; Indefrey & Levelt, 2004; Ito et al., 2016; Pickering & Garrod, 2007). This conclusion, however, has been challenged by reading studies that *do* find evidence of form-based prediction at rates of two words per second (DeLong et al., 2019; Kim & Lai, 2012; Laszlo & Federmeier, 2009) or even four words per second (DeLong et al., 2021). But critically, in all four of these studies, the target word was always highly predictable (unlike the manipulation of predictability in Ito et al., 2016). One possibility is that all of these additional constraints might facilitate prediction—but we postpone further discussion of these findings until the General Discussion.

To the best of our knowledge, there has been no work that directly explores whether form-based prediction persists in a rich discourse context where the primary goal is comprehension. A richer discourse could facilitate prediction by introducing stronger constraints that unfold gradually over time, ensuring that the relevant lexical items are highly active long before the word itself is uttered. Alternatively, natural discourse may well be more variable than typical psycholinguistic stimuli, making prediction more complex and potentially less advantageous. After all, we usually speak because we have something *new* and potentially *unexpected* to convey.

## The Present Study

To explore the degree to which form-based prediction occurs in naturalistic contexts, we used a novel comprehension task called the *Storytime* paradigm (Yacovone et al., 2021; Levari & Snedeker, 2024). Rather than using hundreds of unrelated sentences as stimuli, this paradigm uses coherent, naturally produced stories. These stories can be presented intact in a correlational design that explores

how naturally occurring variation affects the ERP signals at each word (Brennan et al., 2016; Brennan & Hale, 2019; Levari & Snedeker, 2024; Li et al., 2021), or alternatively, we can use these stories as a substrate into which experimental manipulations are spliced (Yacovone et al., 2021).

In the present study, we adopt the latter approach and splice in manipulations that resemble those from the form-based prediction studies described above: Participants will occasionally hear nonwords with varying degrees of similarity to the original word from the story (e.g., *ceke* when the original word is *cake*). Unlike prior studies, however, we did not include a condition in which the nonword shares no features with the target (e.g., *tont* when the original word is *cake*) because we did not want to completely disrupt the flow of the story. Instead, we compared the highly similar nonwords (*ceke*) with nonwords that have a different onset but the same rime (e.g., *vake* when the original word is *cake*). These rime conditions are similar to those in Van Petten et al. (1999). Finally, like Ito et al. (2016), we wanted to investigate form-based prediction in both predictable and unpredictable environments, which we identified using the cloze procedures described below. Thus, our experiment had a factorial design with three word types (*cake*, *ceke*, *vake*) in two predictive contexts (*predictable*, *unpredictable*).

Given the design of our study, we predicted that there will be a reduction in the N400 for similar nonwords (*ceke*) relative to the less-similar nonwords (*vake*) in the most predictable contexts. In the less predictable contexts, there should be no differences between the N400s for the two nonwords. This data pattern would provide strong evidence that form representations were preactivated in the predictable contexts, leading to easier processing of the form-similar nonword after being encountered in the input. Such evidence would also support the idea that form-based prediction *does* occur during naturalistic listening tasks.

In addition to our main prediction, we also had a set of general expectations and secondary predictions about other data patterns that may emerge: First, for non-manipulated baseline words (*cake*), the N400 responses should become smaller as the predictability of the words increase—that is, predictable words should have more *positive* N400 responses than unpredictable ones. This is because the N400 is inversely correlated with the predictability of a word given its context (Kutas et al., 2019; Kutas & Hillyard, 1984). Second, the N400 effects in the more predictable contexts may emerge earlier than those in less predictable ones (e.g., Brothers et al., 2015). Finally, we anticipated that the P600 responses to nonwords would be larger in higher constraint relative to lower constraint contexts (Gunter et al., 2000; Ito et al., 2016; Kuperberg et al., 2020; Vissers et al., 2006).

# Method

## Participants

We recruited 38 native English-speaking adults from the Greater Boston area. Four adults self-reported speaking a non-American dialect of English (British English). All participants provided consent and received two study credits or cash payment for their time. We excluded eight participants from our final analyses following our preregistered exclusion criteria: three for having more than 25% trial loss after data processing, four for poor attention to the task (e.g., falling asleep), and one for researcher error. After these exclusions, we had 30 participants for our final analyses. Participants also completed a modified version of the Language Experience and Proficiency Questionnaire (see Kaushanskaya et al., 2020) in which they reported information about their language background, as well as other demographics such as age (*mean age* = 22 years, *range* = 16–40 years) and gender (*female* = 18, *male* = 10, *nonbinary* = 2).

## Stimuli

The present study used a novel EEG task called the *Storytime* paradigm, which involves taking naturally produced stories and splicing in carefully controlled experimental manipulations (see Yacovone et al., 2021). Specifically, we used an abridged version of a children's book called *Mystery of the Turtle Snatcher* by Kyla Steinkraus as the substrate for our experimental design. We also created a cartoon for this story, which participants watched while listening to the story narration. To preview our design, we selected 180 target words within the story to create a 2 × 3 manipulation of predictability and word type. First, we selected target words with high or low predictability given their preceding story contexts (as determined by a cloze probability task). Predictable target words had higher cloze probabilities, whereas unpredictable target words had lower cloze probabilities. Then, we created three alternative productions for each target word: the original word, a form-similar nonword, or a less-similar rime nonword. This resulted in the six conditions shown below in Table 1. In the remainder of this section, we provide additional details about how these words were selected and how the audio and cartoon stimuli were created.

### *Selecting Our Story Substrate*

We selected *Mystery of the Turtle Snatcher* for two reasons: First, we plan to conduct a parallel experiment with young children. Thus, we needed a story with age-appropriate language and a simple, child-friendly plot. Second, we wanted to use a story that participants were unlikely to be familiar with so that the cloze probability measures would reflect the predictability of the target words given the discourse context rather than prior knowledge of the story itself. Because the original story was too long to present in a single EEG recording session, we created an abridged version by eliminating

**Table 1**
*Example Sentences From the Story*

| Condition | Example sentence |
|---|---|
| High cloze, Baseline | These hairs prove you were at the scene of the *crime* [kraɪm]. |
| High cloze, Similar | These hairs prove you were at the scene of the *crame* [kreɪm]. |
| High cloze, Rime | These hairs prove you were at the scene of the *nime* [naɪm]. |
| Low cloze, Baseline | They are only found in a certain *river* [rɪvər] in Texas. |
| Low cloze, Similar | They are only found in a certain *ruver* [rʌvər] in Texas. |
| Low cloze, Rime | They are only found in a certain *piver* [pɪvər] in Texas. |

some nonessential passages. This version was roughly 30 min when read aloud.

### Assessing the Predictability of Our Story

To find predictable and unpredictable words for our study, we conducted a cloze task (e.g., Taylor, 1953) in which we determined the predictability of *every* word in our story. Specifically, we recruited 541 participants on Amazon Mechanical Turk (https://www.mturk.com) and asked them to complete sentences from the story by guessing each word, one after another (e.g., *The …*, *The cat …*, *The cat was …*, *The cat was hungry*). Participants guessed around 300 words from a single section of the story and read the remainder of the text sentence-by-sentence. Occasionally, participants would see illustrations from the original chapter book. We excluded 91 participants for failing data quality checks, resulting in 450 participants in the final sample. After these exclusions, we had 30 observations for each word in the story, which we used to calculate cloze probabilities.

In the present study, we define a word's cloze probability as the proportion of trials in which participants correctly guessed that word—for example, if 30 participants provided a guess for the word and 27 participants guessed it correctly, that word would have a cloze probability of 90% (i.e., 27 of 30) given its preceding context. This approach, however, is slightly different from approaches that use cloze tasks to characterize whether participants converge on *any* word given the context. One could imagine a situation in which the context is highly constraining but leads participants to guess a word that was not actually used in the story itself. For example, the sentence "I like my coffee with cream and *cinnamon*" is apt to be completed with *sugar* instead of *cinnamon*. Thus, cinnamon would be a low cloze word in a highly constraining context. In the present study, however, cloze and constraint were categorically linked, as the degree of constraint for high cloze items (*mean constraint* = 81.2%, *SE* = 1.8%) and low cloze items (*mean constraint* = 44%, *SE* = 1.8%) significantly differed ($b = -0.37$, $SE = 0.03$, $t = -14.21$, $p < .001$).

### Selecting the Target Words

To select our target words, we first calculated the cloze probabilities for all common nouns in the story. Then, we sorted these nouns from highest to lowest cloze and removed all nouns that started with vowel sounds (because we could only create the rime condition in a consistent way if the target word began with a consonant). We also removed words from the list if they appeared in the same sentence as another noun that had a more optimal cloze probability (i.e., higher for predictable targets or lower for unpredictable targets). Next, we removed the additional tokens of nouns that occurred more than three times in the story (e.g., turtle) to ensure that participants never heard the same manipulation more than once. In choosing which tokens of a given noun to keep, we preferentially selected those with the most extreme cloze values (either high or low). Finally, we had to remove nouns that could not be changed into nonwords following the process described below. For example, *pan* could not be turned into a form-similar nonword by changing the first vowel because all possible candidates are in fact real words (e.g., pawn, pain, pin, pen). After all of these exclusions, we selected the top 90 words for the high cloze targets and the bottom 90 words for the low cloze targets. The high cloze targets had an average cloze probability of 81.2% (*SD* = 14.2%, *range* = 53%–100%), and the low cloze targets had an average of 7.2% (*SD* = 13.8%, *range* = 0%–50%).[3] We also characterized the lexical frequency of our target words and their duration in milliseconds. To do this, we collected standardized word frequencies (per million words) from the $SUBTLEX_{US}$ corpus (Brysbaert & New, 2009), which contains roughly 51 million words from American English subtitles between 1990 and 2007. Note, however, two low cloze target words did not appear in the $SUBTLEX_{US}$ corpus (*whiteboard*, *hatchlings*). There were no significant differences between the frequency of the high cloze (*mean frequency* = 200.37, *SD* = 352.66) and low cloze (*mean frequency* = 123.65, *SD* = 361.23) conditions, $t(175.62) = 1.43$, $p = .15$. For word duration, we ascertained the length of the target words in milliseconds using the Gentle forced aligner (Ochshorn & Hawkins, 2016). There were no significant differences between the duration in ms of the high cloze (*mean duration* = 508.9, *SD* = 0.14) and the low cloze (*mean duration* = 509.4, *SD* = 0.28) baseline conditions, $t(127.97) = -0.02$, $p = .99$.

### Creating the Nonword Violations

We created three conditions for each of the 180 final target words: the original word, the form-similar nonword, and the less-similar rime nonword. The original word was the intended target word from the story. To create the form-similar nonword, we changed the first vowel sound of each target word, ensuring that the vowel change did not result in another word of English (e.g., *nap* [næp] became *nupe* [nup] and not *nip* [nɪp]). Some of these changes, however, may have resulted in extremely low frequency words (e.g., beal) or words from non-American dialects of English (e.g., lud, mooth). To create the rime nonwords, we changed the first consonant of each target word (e.g., *cage* became *nage*). We wanted the consonant change to be maximally different—so, we implemented changes on three dimensions: place of articulation, manner of articulation, and voicing. For example, the [*k*] in *cage* is a voiceless, velar stop, whereas the [*n*] in *nage* is a voiced, alveolar nasal.

We also used the *Irvine Phonotactic Online Dictionary* (Vaden et al., 2009) to calculate the unstressed phonological neighborhood density for all baseline targets and their nonword manipulations. We then tested for statistical differences between all of the critical pairwise comparisons. Importantly, we did not find any significant differences between conditions with respect to their phonological neighborhood densities (see our analysis code on the Open Science Framework [OSF]: https://osf.io/upjc4/).

[3] We also conducted an auditory cloze task in which a new set of 45 English-speaking adults watched the cartoon narration (further described in Stimuli section) and then guessed a subset of our target words. Thus, we were able to collect and compare spoken and written cloze probabilities for all 180 target words in the present study. We observed a strong correlation between the cloze values ascertained from the written and spoken cloze tasks, which was statistically confirmed via a Spearman's rank order correlation, $r_s(178) = .78$, $p < .001$. Overall, the target words had slightly higher cloze value scores in the spoken cartoon task (*mean cloze* = 53.2%, *median cloze* = 63.6%, *SD* = 39.1%) than in the written task (*mean cloze* = 44.2%, *median cloze* = 51.7%, *SD* = 39.7%), as indicated by a paired-samples Wilcoxon signed rank test, $z = -4.84$, $p < .001$, 95% CI [−0.12, −0.05]. For additional details and group comparisons, see our additional online materials on OSF (https://osf.io/upjc4/).

### *Creating the Spliced Recordings and Cartoon Stimulus*

After constructing all 540 target sentences (180 total target words × 3 word types), the first author, who is a native English speaker, recorded the materials. First, he recorded the 30-min story in its entirety. We used this story as the substrate into which we spliced our manipulations. Next, he recorded the 540 target sentences in isolation, making sure to replicate (to the best of his ability) the intonational and prosodic contours of the original recording. We then extracted the critical target words from these isolated sentence recordings and spliced them into the base recording. In some sentences, we needed to extract a few words before and/or after the target word to avoid issues with coarticulation and prosody when splicing.

This splicing procedure ensured two properties of our stimuli: First, all conditions, regardless of being a nonword or the intended word, had been spliced in from a different audio file; and second, the auditory context before and after the target word (or region) was held constant across experimental lists. We constructed three experimental lists using a pseudo-Latin Square design, ensuring that no two target words from the same condition appeared back-to-back. We also ensured that all repeated target words appeared in different conditions in each list. To meet these criteria, we needed to have a slight difference in the number of observations per cell in each list. For example, one list had the following distribution: for high cloze targets, 27 baseline words (*cake*), 30 form-similar nonwords (*ceke*), and 33 rime nonwords (*vake*); for low cloze targets, 33 baseline words, 27 form-similar nonwords, and 30 rime nonwords.

To encourage participants to pay attention to the story, we created a cartoon to accompany it. Examples of stills from the cartoon can be found in Figure 1. The cartoon was created using Vyond software (https://Vyond.com), and the first author was unaware of which target words would be selected from the story at the time of making the cartoon. Thus, we did not design the cartoon to alter the predictability of the target words.

In the *Storytime* paradigm, we do not construct specific sentences to serve as fillers. Instead, we rely on the sentences in the text that have not been manipulated to serve the functions of fillers (e.g., making the manipulations less predictable and reinforcing the expectation that most sentences do not have errors). Our story had around 500 sentences, and 180 of them contained target words. Thus, there was a ratio of roughly 2:1 between fillers and targets. Moreover, only two thirds of the target sentences contained a violation, making the ratio of correct-to-incorrect sentences 3.5:1. The stimulus onset asynchrony, which represents the amount of time from the onset of one target to the onset of the next, was 10.25 s on average (*range* = 1.33–47.81, *SD* = 8.62). To estimate the speech rate, we divided the total number of words in our story by the total amount of time spent speaking (i.e., the total phonation time). Specifically, we had 4,634 words in our story and an average phonation time of 1,427.32 s, as calculated by a PRAAT script from de Jong and Wempe (2009). We found that, on average, our three story versions were produced with an average speech rate of 3.25 words/s.

## Procedure

### *Experimental Setup*

Participants listened to the story while watching the cartoon in a single 30-min EEG recording session. The cartoon was presented using PsychoPy (Peirce et al., 2019). Participants sat roughly 100

**Figure 1**
*Stills Taken From our Cartoon Stimulus*

*Note.* This cartoon was presented alongside the story narration to promote attention to and understanding of the discourse context. The full cartoon video is available on our OSF page (https://osf.io/upjc4/). See the online article for the color version of this figure.

cm from a TV monitor, and they were encouraged to minimize movement and to keep their faces relaxed.

### EEG Recording

Participants were fitted with an electrode cap (actiCAP SnapCap) containing 31 active Ag/AgCl electrodes that were connected to the EEG equipment, Brainvision's actiCHamp Standard 64 System. Two external mastoid electrodes (TP9 and TP10) were placed directly behind participants' ears. The EEG data were recorded at a sampling rate of 500 Hz using Brainvision's Recorder (BrainVision Recorder, Version 1.23.0001, Brain Products GmbH, Gilching, Germany). On average, electrode impedances were kept below 20 kΩ. During recording, the ground electrode was FPz, and the reference electrode was FP1.

### Data Preprocessing Steps and Data Exclusion Criteria

We used the EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes in MATLAB (The MathWorks Inc., 2022) to preprocess the EEG data. Our procedure for preprocessing is as follows: First, we re-referenced the data to the average of the left and right mastoid electrodes. Then, we applied a high-pass filter of 0.1 Hz and downsampled the data from 500 to 200 Hz. Next, we extracted 2,000 ms epochs from the continuous data between −500 and 1,500 ms relative to stimulus onset (without baselining). We then conducted an independent component analysis with these epochs in order to identify and correct EEG artifacts (including blinks and horizontal eye movements). Independent component analysis components were classified using ICLabel (Pion-Tonachini et al., 2019) and then corrected if they received at least 75% probability of belonging to the following artifact groups: muscle, eye, heart, line noise, or channel noise. Then, we extracted our target epochs from −200 to 1,500 ms with a 200 ms pre-stimulus baseline. These epochs were subjected to an automatic artifact rejection procedure, which removed trials with voltages exceeding ±100 μV. If necessary, electrode channels with greater than 5% trial loss were interpolated; however, we never interpolated more than 10 channels for a single participant. In fact, we only interpolated 19 channels in total and, on average, less than one channel per participant (*range* = 0–8 interpolated channels per participant). Of these interpolated channels, only two were used in our final analyses. Finally, we applied a low-pass filter of 30 Hz.

All participants with more than 25% of their trials rejected after these cleaning procedures were excluded, and their data were replaced. On average, we rejected 4.75% of participants' total trials (*SD* = 5.29%, *range* = 0%–22.8%), and we only had to reject eight participants: three for data loss, four for failing to attend to the task (e.g., falling asleep), and one for researcher error.

## Statistical Analyses

### Determining Our Regions of Interest

To determine our spatial regions of interest (ROIs), we relied on information from the prior literature and our own pilot data. For N400 effects, we selected a centroparietal ROI with eight electrodes: Cz, C3, C4, CP1, CP2, Pz, P3, and P4. For P600 effects, we selected a parietal ROI with three electrodes: Pz, P3, and P4. To determine our temporal ROIs, we used a collapsed localizer technique in which all the conditions being compared were collapsed into one grand average waveform, which was then used to determine the ROI (for discussion, see Luck & Gaspelin, 2017). Researchers vary in how they use the grand average—some rely on visual inspection, some conduct cluster-mass permutation tests on these averages, and some select a time window based on the highest peak within a window (Luck, 2014; Luck & Gaspelin, 2017). We used this last approach. Specifically, we used a 200 ms time window centered on the most negative peak (for N400 effects) and the most positive peak (for P600 effects) in the grand average waveforms.

For analyses that directly compared high cloze and low cloze conditions, we defined two temporal ROIs, creating a grand average for each cloze condition but still collapsing across word type. This is because prior work has shown that predictability can influence the timing of ERP effects (Kutas & Federmeier, 2011; Swaab et al., 2012). For example, N400 effects can emerge earlier for words in predictable contexts relative to unpredictable ones (Brothers et al., 2015).

### Linear Mixed-Effects Model Specifications

All linear mixed-effects models were implemented using the *lme4* (*v*. 1.1-31, Bates et al., 2014) and *afex* (*v*.1.2-1, Singmann et al., 2023) packages in R (*v*. 4.2.2, R Core Team, 2022). All pairwise comparisons were implemented using the *emmeans* package (*v*. 1.8.4-1, Lenth, 2024). We initially fit our models with the maximal random effects structures justified by our data. If we encountered convergence issues, we simplified the random effects structure until the models properly converged (Baayen et al., 2008; Barr, 2021). Most issues were resolved by constraining the covariance parameters for the random effects to zero (i.e., removing the correlations between them). But, if this step did not resolve the issues, we began to incrementally drop random slopes (while trying to preserve the slopes for the highest order effects of interest) until the models converged. All effects with an absolute value of *t* greater than 2 are considered significant (Gelman & Hill, 2006). We follow this convention due to the uncertainty in the field about how to best calculate the appropriate degrees of freedom in linear mixed-effects models (Baayen et al., 2008). But for the sake of completeness, we also report the *p* values as calculated by the *lmerTest* package (*v*. 3.1-3, Kuznetsova et al., 2017). Note that in all models, both methods of evaluating significance arrived at the same conclusions.

### Outline of Our Analyses

For the present study, we preregistered a set of primary and secondary hypotheses. In the Results and Discussion section, we report the findings from five linear mixed-effects models that aimed to test those hypotheses.[4] First, we tested whether the N400s to form-similar nonwords (*ceke*) were reduced relative to less-similar nonwords (*vake*) in high but not low cloze contexts.

[4] The analyses reported in the Results and Discussion section slightly differ from those in our preregistration due to recommendations that we received during an external review of this work. The main difference between them is that the preregistered analyses modeled by-participant or by-item ERP averages whereas the reported analyses modeled trial-level ERPs. A full comparison of these two analytical approaches can be found in the additional online materials on OSF (https://osf.io/upjc4/). Critically, both approaches resulted in the same patterns of findings.

Second, we tested whether our P600 effects were sensitive to predictability and word type. Third, we investigated our secondary hypotheses about item-level differences in our ERP effects. To do this, we first analyzed how our N400 responses changed as a function of a word's predictability, i.e., did baseline N400s show an inverse linear relationship with cloze probability? Then, we conducted a set of parallel analyses for our P600 effects. The results of these analyses are briefly discussed in the following section, as well as more thoroughly in the General Discussion.

## Transparency and Openness

### *Data, Analysis Code, and Research Materials*

The preregistration for this study, as well as all of the data, analysis scripts, and research materials are available on our OSF page (https://osf.io/upjc4/; Yacovone et al., 2024). We analyzed our data using the R statistical computing environment (*v*. 4.2.2, R Core Team, 2022). Data visualizations were created using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes in MATLAB (The MathWorks Inc., 2022) and the *ggplot* package in R (*v*. 3.4.2, Wickham, 2016).

### *Sample Size Calculation and Power Analysis*

A priori power analyses were conducted using the *mixedpower* package in R (Kumle et al., 2021; R Core Team, 2022). This package uses a simulation-based approach to estimate the power for each effect of interest across a range of sample sizes. To determine our optimal sample size, we first collected data from 30 pilot participants. Then, we implemented a set of mixed-effects models outlined in our preregistration. We found that the subtlest, reliable effects in our pilot data were the interaction terms. Thus, we wanted to determine the number of participants needed to achieve at least 80% power for all relevant interaction effects with $\alpha = .05$.

For ease of computation, we dropped the random slopes in our simulations. In addition, we did not want model convergence issues to contribute inaccurate model estimates into the simulated distribution of effect sizes. To account for dropping slopes (and to be more conservative with our power calculations), we also implemented a smallest effect sizes of interest approach by reducing our observed effect sizes by 20% and then calculating the sample size necessary to achieve 80% power for these smallest effect sizes of interests (see Kumle et al., 2021). Results indicated that we needed a final sample size of 30 participants.

## Results and Discussion

In Figure 2, we present the grand average waveforms from all centroparietal electrodes (Cz, C3, C4, CP1, CP2, Pz, P3, P4) for each word type in both cloze conditions. These waveforms were calculated by first collapsing across items to get six waveforms for each participant and then collapsing across those by-participant averages. Visual inspection shows robust N400 and P600 effects for all violation conditions, regardless of predictability. Critically, in the high cloze condition, the N400 effect is reduced for the form-similar nonwords (*ceke*) relative to the less-similar rime nonwords (*vake*) across all electrodes (see Figure 2, top panel). In contrast, in the low cloze condition, there is no evidence of a reduction for the form-similar nonwords (see Figure 2, bottom panel). Finally, for the baseline conditions (*cake*), the N400 responses are smaller for high cloze words than for low cloze words, as predicted.

In Figure 3, we present the topographic maps for the effects of our form-similar and rime manipulations. In high cloze contexts, the reduced N400 for form-similar nonwords can be seen as a weaker negative effect in the 400 ms and 600 ms time windows relative to the rime condition (see Figure 3, top panel). Both violations in high cloze conditions also produced large P600s that were similar in magnitude and latency (see 1,000 ms and 1,200 ms). For low cloze contexts, both violations seemed to elicit similar N400 and P600 effects (see Figure 3, bottom panel). In addition, the N400 effects for the high cloze conditions emerged slightly earlier than those in the low cloze conditions, as predicted.

## Are N400 Effects Reduced for Form-Similar Errors in Predictable Contexts?

To demonstrate form-based prediction, we would need to show a significant reduction in N400 effects for form-similar nonwords (relative to the less-similar rime nonwords), but only when participants are actually able to predict the original word from the story. Figure 2 provides initial evidence that form-based prediction is occurring, as *ceke* evoked smaller N400 responses than *vake* in high cloze but not low cloze conditions. To confirm that these differences are statistically reliable, we first calculated mean N400 responses by averaging the amplitudes from our preregistered centroparietal electrodes in the time windows identified by the localizers. These localized time windows were 420–620 ms and 430–630 ms for high and low cloze words, respectively. We then modeled these mean N400 responses using a linear mixed-effects model. This model had fixed effects of word type (*cake*, *ceke*, *vake*) and predictability (*high cloze*, *low cloze*), as well as their interaction. For word type, we used contrast coding to test for successive differences between *cake* and *ceke* and then between *ceke* and *vake*. For predictability, we used contrast coding to test for differences between high and low cloze conditions (*high cloze* = −.5, *low cloze* = .5).[5] The remaining pairwise comparisons were tested (and corrected for multiplicity) using the *emmeans* package (Lenth, 2024). Finally, this model had random intercepts and maximal slopes for participants and items. To reach convergence, we constrained the covariance parameters for the random effects to zero (see Baayen et al., 2008; Barr, 2021).

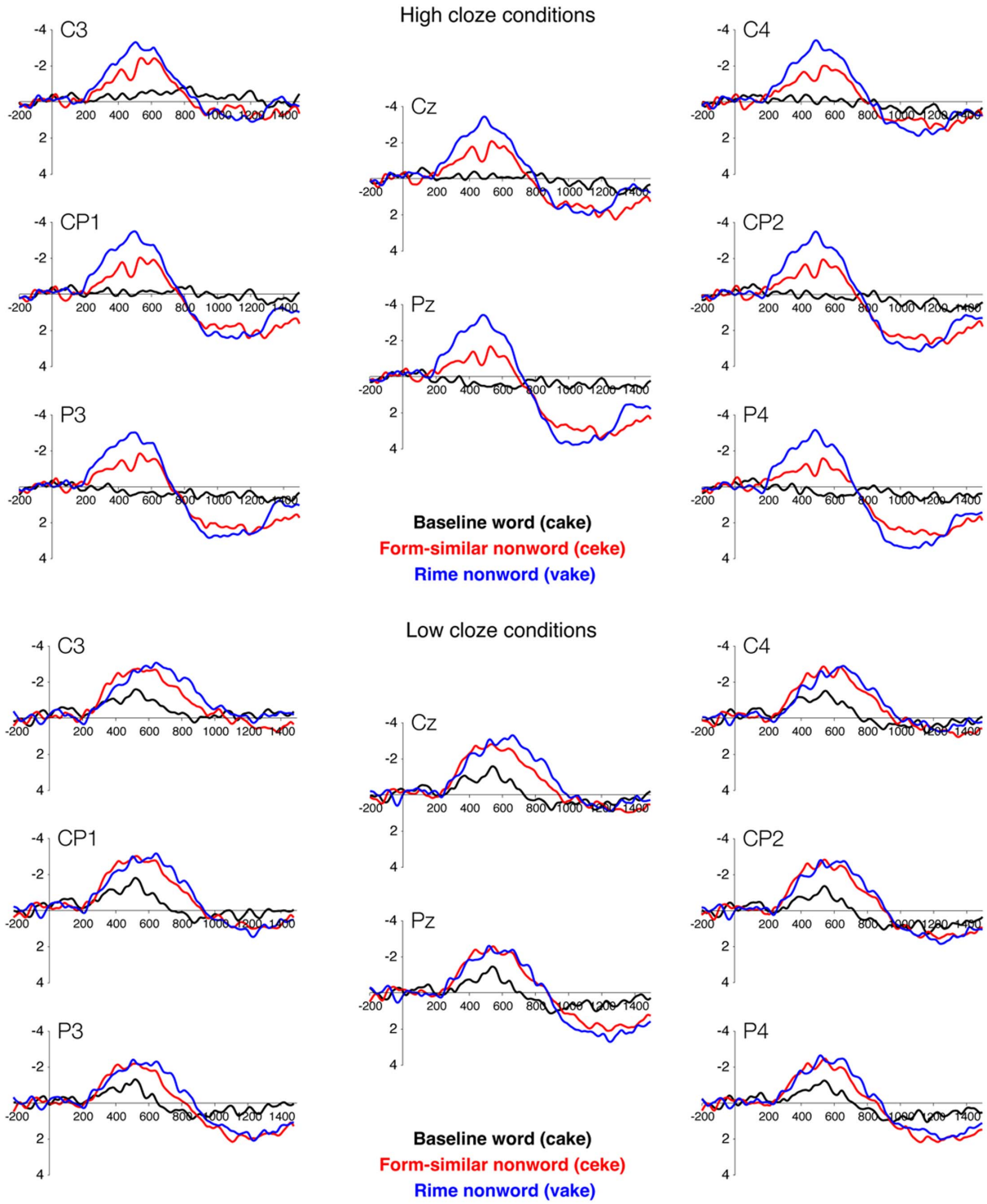
Results indicated a significant main effect for the word type contrast between *cake* and *ceke* ($b = -1.61$, $SE = .40$, $t = -4.01$, $p < .001$). There were no main effects for the contrast between *ceke* and *vake* ($b = -0.67$, $SE = .36$, $t = -1.89$, $p = .067$) nor predictability ($b = 0.51$, $SE = .42$, $t = 1.22$, $p = .23$). There was, however, a significant interaction between word type and predictability, but only for the contrast between the two nonwords, *ceke* and *vake* ($b = -1.52$, $SE = .71$, $t = -2.15$, $p = .033$).[6] To unpack this

[5] We also implemented a parallel model with *z*-scored control predictors of lexical frequency, word duration, and phonological neighborhood density. Critically, the patterns of significance remained the same.

[6] We also conducted a parallel analysis using a more traditional time window of 300–500 ms for both high and low cloze conditions, as well as controlling for lexical factors such as frequency, word duration, and phonological neighborhood density. Both analyses resulted in the same pattern of significance—and moreover, the model estimate and *t* value for the critical interaction between *ceke* and *vake* increased ($b = -1.92$, $SE = 0.66$, $t = -2.91$, $p = .004$).

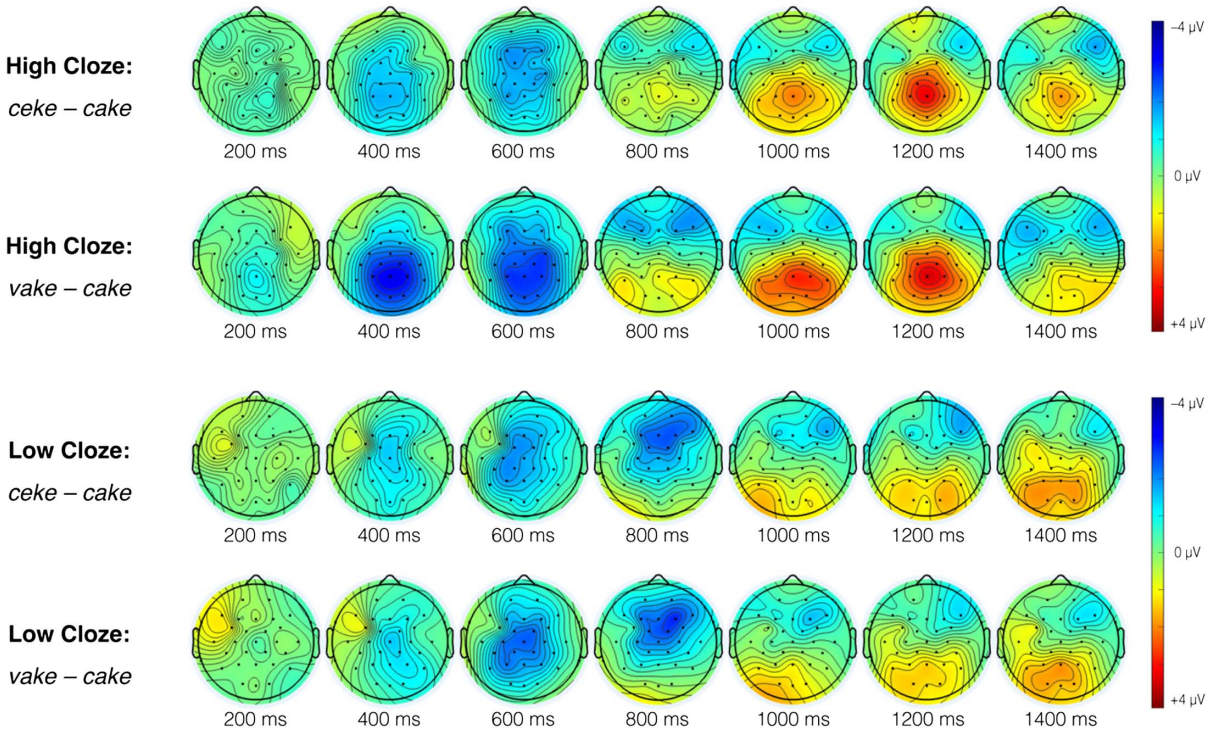
**Figure 2**

*Grand Average ERP Waveforms by Word Type and Cloze Condition*




*Note*. The grand average waveforms (μV) for the centroparietal electrodes of interest are presented for both high cloze (top panel) and low cloze (bottom panel) conditions. The black lines represent the baseline condition (*cake*), while the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions, respectively. All waveforms were subjected to an additional low-pass filter of 15 Hz for plotting purposes. ERP = event-related potential. See the online article for the color version of this figure.

**Figure 3**
*Topographic Maps of the ERP Effects by Cloze Condition*




*Note*. These topographic maps depict the isolated effects of the form-similar (*ceke*) and rime (*vake*) nonwords in both predictable (high cloze, top panel) and unpredictable (low cloze, bottom panel) contexts. These effects are calculated by subtracting the baseline ERP activity from the activity evoked by each violation. ERP = event-related potential. See the online article for the color version of this figure.

interaction, we conducted planned pairwise comparisons within predictability conditions, which revealed that the N400 responses for *ceke* were significantly smaller than those for *vake* in the high cloze condition ($b = 1.43$, $SE = .50$, $t = 2.85$, Tukey-adjusted $p = .016$) but not in the low cloze condition ($b = -0.08$, $SE = .50$, $t = -0.17$, Tukey-adjusted $p = .98$). All of the remaining pairwise comparisons within predictability groups were significant.

These results confirmed that there were robust N400 effects for all error types across both high and low cloze words, however, the N400 effects for *ceke* were only significantly smaller than those for *vake* in high cloze environments. In contrast, there was no reduction in the N400 effects for *ceke* in the low cloze environments, which suggests that participants were unable to predict the form of the original word in those less predictable contexts.

## Are P600 Effects Different Across Error Type and Predictability?

To determine whether P600 effects are sensitive to error type and predictability, we followed a similar procedure to the one outlined above: First, we calculated mean P600 amplitudes from our preregistered parietal electrodes in the time windows from the localizers. Those time windows were 1,085–1,285 ms and 1,160–1,360 ms for high and low cloze, respectively. Then, we implemented a linear mixed-effects model using the same fixed effects of word type (*cake*, *ceke*, *vake*), predictability (*high cloze*, *low cloze*), and their interaction. For word type, we used contrast coding to test for differences between *cake* and *ceke* and then between *cake* and *vake*. Note, this last comparison is different from the one used in the N400 analysis above, reflecting a difference in the hypothesis being tested. For predictability, we again compared the two cloze conditions (*high cloze* = −.5, *low cloze* = .5). Finally, this model had random intercepts and maximal slopes for participants and items. To reach convergence, we constrained the covariance parameters for the random effects to zero.

Results indicated main effects for both word type contrasts, confirming the robust P600 effects seen in Figures 2 and 3 for both *ceke* ($b = 1.81$, $SE = .44$, $t = 4.15$, $p < .001$) and *vake* conditions ($b = 2.08$, $SE = .41$, $t = 5.07$, $p < .001$). There were no main effects of predictability ($b = -0.58$, $SE = .56$, $t = -1.04$, $p = .30$) nor any interactions, suggesting that the observed differences in magnitude across high and low cloze conditions were not statistically reliable. We return to this point in the General Discussion.

## Do N400s and P600s Vary Continuously With Word-Level Predictability?

In addition to our primary analyses, we also preregistered a set of secondary analyses that explore whether word-level predictability modulates our observed effects. In the first analysis, we sought to replicate the finding that N400 responses to nonmanipulated words (*cake*) are inversely correlated with the predictability of that word

given its particular context (Kutas et al., 2019; Kutas & Hillyard, 1984). In the second analysis, we explored how the N400 responses to our violations changed as a function of the original word's predictability. For example, does the N400 reduction for *ceke* increase linearly with predictability (similar to the N400 reduction for *cake*)? In the final analysis, we investigated parallel questions about the relationship between P600s and word-level predictability.

Before addressing these questions, we first visualized how our grand average waveforms changed across cloze probability by categorizing items into one of three bins: lowest cloze, middle cloze, and highest cloze (see Figure 4). To create these waveforms, we first averaged across participants to get three waveforms per item. Then, we averaged across these by-item averages within the same cloze bin. Figure 4 provides some insight into our three secondary hypotheses above: First, with respect to the baseline conditions, the N400 responses become less negative across the cloze bins. Second, the reduction in N400 responses for *ceke* becomes more robust as predictability increases. Interestingly, the size of the N400 response to *vake* seems to slightly increase across predictability, moving in the opposite direction of the form-similar effect. Finally, the magnitude of the P600 effects appears to be similar across error types and across cloze bins, supporting our findings from the primary analyses above.

### *Do Baseline N400s Become Smaller as Predictability Increases?*

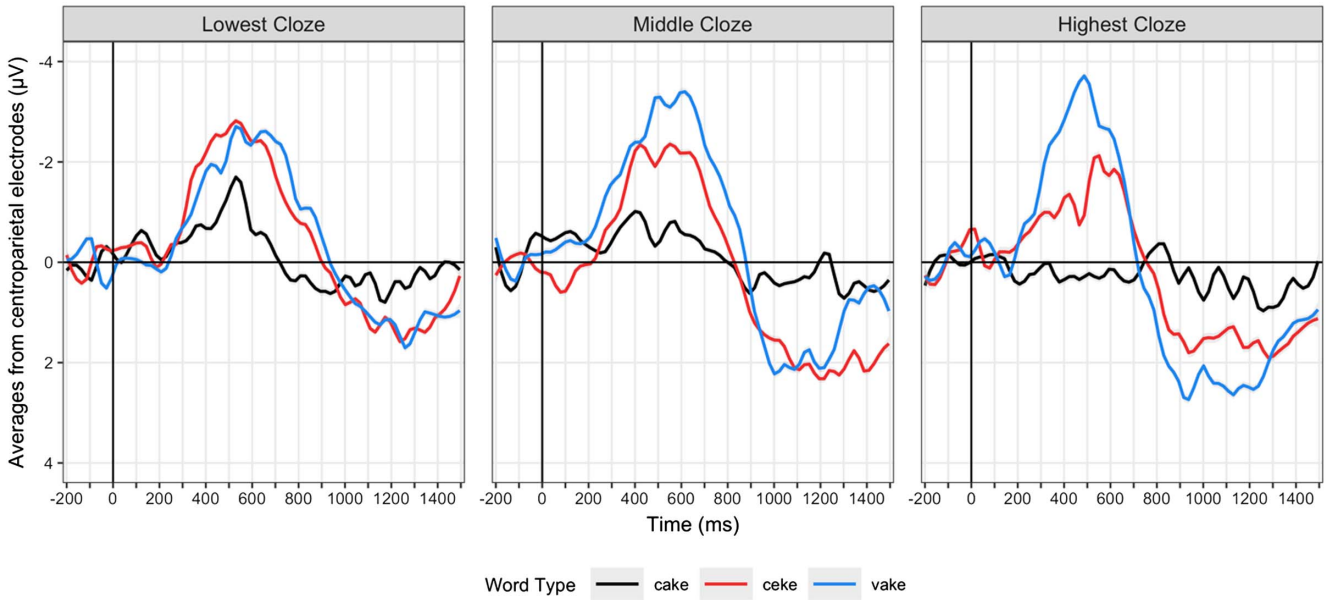
It is well-documented that the N400 responses to nonmanipulated baseline words become smaller (or more positive) as a function of the word's predictability in its particular context (Kutas et al., 2019; Kutas & Hillyard, 1984). Our design allows us to investigate whether this data pattern is present in naturalistic contexts and with stimuli that contain frequent violations. To do this, we modeled the trial-level data from above using a linear mixed-effects model with a single continuous fixed effect of cloze probability. For random effects, the model had random intercepts for both participants and items (as justified by the data). Results confirmed a significant effect of cloze probability such that the N400 amplitude for baseline words becomes more positive as predictability increases ($b = 1.78$, $SE = .76$, $t = 2.33$, $p = .021$). Figure 5 below shows the change in N400 amplitudes across cloze probability values for all word types.

### *How Do the N400s to Both Error Types Change Across Predictability?*

Next, we wanted to investigate how predictability modulates the size of the N400 for the form-similar (*ceke*) and rime (*vake*) nonwords. To do this, we originally planned to isolate the effects by calculating difference waves, i.e., by subtracting the baseline (*cake*) from both error conditions as in Figure 3. However, in constructing these difference waves, we realized that the substantial effect of cloze on the baseline condition makes it difficult to interpret this analysis in isolation. So, although this approach deviates slightly from our preregistration, we thought it would be informative to test all word types in a single linear mixed-effects model to quantify their similarity. Specifically, we implemented a new model with fixed effects of word type (*cake, ceke, vake*), word-level cloze probability (continuous 0%–100%, mean-centered at 44%), and their interaction. As in the primary models, we used contrast coding to test for successive differences between *cake* and *ceke* and then between *ceke*

**Figure 4**

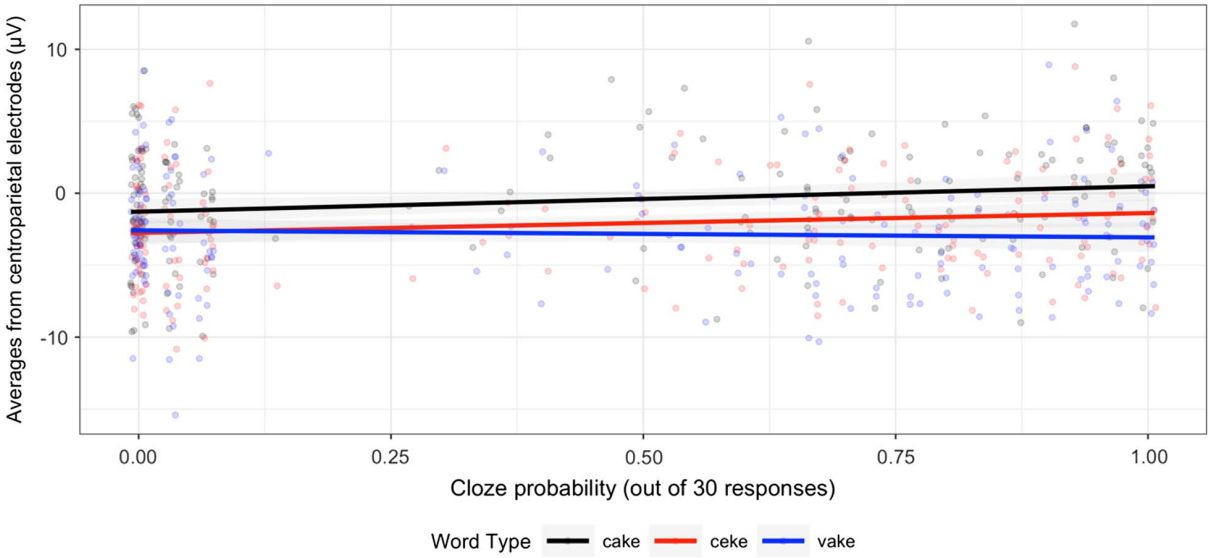
*Visualization of the ERP Effects Across Cloze Probability Bins*




*Note*. The grand average waveforms (μV) are plotted in three cloze probability bins, increasing in predictability from left to right. The black lines represent the baseline condition (*cake*), and the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions, respectively. These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques. ERP = event-related potential. See the online article for the color version of this figure.

**Figure 5**

*Visualization of N400 Amplitudes Across Cloze Probability for Each Word Type*




*Note*. Each line represents the linear trend in N400 amplitudes as cloze probability increases from left to right. The black line represents the baseline condition (cake), and the red and blue lines represent the form-similar (ceke) and rime (vake) conditions, respectively. Each dot represents the average N400 amplitude (μV) for each item (180 Items × 3 Word Types = 540 observations). We collapsed across participants for plotting purposes, but we preserved the participant and item-level structures in our statistical analyses. See the online article for the color version of this figure.

and *vake*. For random effects, the model converged with random intercepts and maximal slopes for participants and items.

Results indicated a significant effect between *cake* and *ceke* ($b = -1.61$, $SE = .39$, $t = -4.10$, $p < .001$), as well as a significant interaction between word type and cloze probability for the contrast between *ceke* and *vake* ($b = -1.81$, $SE = .90$, $t = -2.02$, $p = .045$). To follow up on this interaction, we estimated the slopes for each word type individually using the *emtrends* function (see Lenth, 2024). For *cake*, we found identical results to the analysis above. For the two error conditions, the slopes did not reach statistical significance—although *ceke* was estimated to become slightly more positive across cloze ($b = 1.34$, $SE = .71$, $t = 1.89$, $p = .06$), whereas *vake* was estimated to become slightly more negative ($b = -0.47$, $SE = .76$, $t = -0.62$, $p = .54$). These findings tentatively suggest that *cake* and *ceke* experience similar degrees of N400 reduction as word-level predictability increases. Figure 5 shows the relative changes in N400 amplitudes for each word type across cloze probability.

### *How Do the P600s to Both Error Types Change Across Predictability?*

We conducted parallel analyses on our P600 effects; however, none of these analyses revealed any significant insights. Specifically, across all of the analyses above, we found robust P600 effects for both violations, with no differences in magnitude between them. Moreover, we found no effects of predictability nor any interactions between predictability and word type. But, as we mentioned earlier, the cloze probability of a target word is not necessarily indicative of the constraint of the target sentence. For example, in the introduction, we demonstrated this point using the sentence "I like my coffee with cream and *cinnamon*." This particular sentence is apt to be completed with *sugar* instead of *cinnamon*, meaning that *cinnamon* is a low cloze completion for this highly constraining context. In the General Discussion, we present an exploratory analysis of P600 effects in high constraint contexts with low cloze target words. To foreshadow these findings, we show robust P600 effects for all high constraint contexts, regardless of the cloze probability of the target word from the story. These findings are consistent with the proposal that P600s reflect the recognition of a conflict or failure to incorporate the bottom-up input into the comprehenders' higher level interpretation of the unfolding context—but only when the context is sufficiently constraining such that higher level interpretations are actually built (e.g., Brothers et al., 2022; Ito et al., 2016; Kuperberg, 2007; van de Meerendonk et al., 2010; van Herten et al., 2005; Vissers et al., 2006).

## General Discussion

In the present study, we directly explored whether adults predict the form of upcoming words while listening to a rich, naturalistic discourse. To do this, we manipulated a set of target words within a children's story such that participants would either hear the original word from the story (*cake*), a form-similar nonword (*ceke* instead of *cake*), or a less-similar rime nonword (*vake* instead of *cake*). In highly predictable contexts, we found that form-similar nonwords (*ceke*) elicited smaller N400 responses than rime nonwords (*vake*). This finding suggests that participants had predicted the form of the intended word in high cloze conditions, resulting in facilitated processing of the form-similar nonword. In less predictable contexts, both kinds of nonwords elicited similarly sized N400 responses. To the best of our knowledge, these findings are the first to demonstrate that form-based prediction occurs not only in tightly controlled

experimental settings but also in contexts that better resemble everyday comprehension.

In the remainder of this General Discussion, we will do four things: First, we reconcile our findings with the prior literature that suggests form-based prediction should *not* readily occur in naturalistic comprehension. Second, we explore how our work relates to prior studies on phonological mismatch effects in spoken language comprehension. Third, we further examine our P600 effects and integrate our findings with the broader literature on these posterior positivities. Finally, we discuss several open questions about form-based prediction and outline potential avenues for future research on this phenomenon using the *Storytime* paradigm.

## Should Form-Based Prediction Be Expected During Naturalistic Comprehension?

There is an ongoing debate in the literature about the limits of form-based prediction and its role in everyday comprehension. Some researchers argue that, while form-based prediction can occur in certain experimental contexts, it is unlikely to occur in most naturalistic settings (Freunberger & Roehm, 2016; Ito et al., 2016; 2017b). Specifically, there are two features that are unusual about the experimental contexts in which form-based prediction has been studied:

First, the manipulations that are used in such studies can help create a context in which making predictions about an upcoming word is unusually useful for the comprehender. If all the target sentences contain a highly predictable word, and that word is the one that is manipulated or replaced, it would make sense to try to anticipate these words (rather than passively process them) in order to reconstruct the intended (or original) meaning of the utterance. Second, in form-based prediction studies, the target sentences are often presented at slower-than-natural rates, which may provide comprehenders with more time to process and use the unfolding words to generate predictions (for discussion, see Ito et al., 2016; but see DeLong et al., 2021). In the sections below, we ask whether these two features are also true of the stimuli in the present study. To foreshadow our results, we do not find any evidence that our story is particularly slow or predictable. Thus, we end with a proposal about why form-based prediction was possible in this rich but highly variable naturalistic context.

### *How Predictable Was Our Story?*

Prior work has found that comprehenders engage in more predictive processing when they are in highly predictable contexts (e.g., Brothers et al., 2015; Lau et al., 2013). For example, in a study by Lau et al. (2013), participants simply read pairs of words and indicated when they saw the name of an animal. The authors manipulated these word pairs to either be semantically related to one another (e.g., salt–PEPPER) or not related at all (e.g., salt–UNCLE). They also manipulated the overall proportion of trials in which the word pairs were related: In one experimental block, 50% of the word pairs were related, so that activating close semantic associates would be helpful about half of the time. In the other block, only 10% of the word pairs were related, and thus, participants who activated semantic associates would be unlikely to correctly predict the second word given the first. As expected, the authors observed a reduction in the N400 to words that were preceded by semantically related words (e.g., salt–PEPPER) relative to words preceded by unrelated words (e.g., salt–UNCLE). Critically, however, this reduction was greater in the experimental block with a higher proportion of related trials relative to the block with a lower proportion, demonstrating that preactivation based on semantic association is greater when semantic predictability is high.

In the present study, we explored comprehension in the context of a written narrative that was read aloud. Like most written narratives, this story had been carefully crafted and edited with the goal of ensuring that it would be easily understood. Edited language like this is unnatural and yet pervasive—unnatural in the sense that it is quite different from the language environment that existed for most of humans' evolutionary history, and pervasive in the sense that it now makes up a sizeable portion of our participants' linguistic input on any given day (e.g., podcasts, movies, news broadcasts, articles, novels, and many social media posts). One might wonder if, on average, edited text is more predictable than spontaneously produced language—especially when the edited text is intended for children, like the story used in the present study.

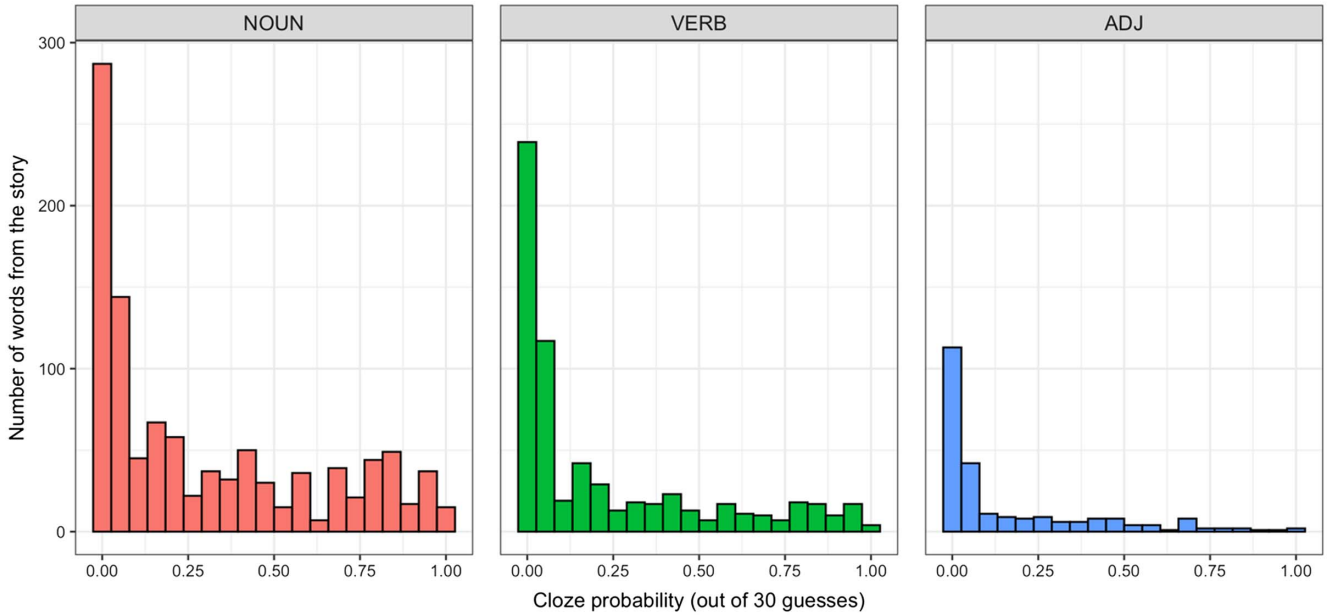
Fully exploring this question goes beyond the scope of the present study. Nevertheless, to better understand the scope and generalizability of our findings, we revisited our cloze data to characterize the overall predictability of our story. In our original cloze task, each person would read the story up until they had to guess, and then they would guess each word in their small section, word-by-word (e.g., *The …, The cat …, The cat was …, The cat was hungry*). Participants saw the correct answer on the screen after making each guess. For example, a person might have guessed that a sentence started with *The*. After guessing, they learn that it started with *She*, and so they update their expectations about how the sentence will unfold and then guess the word *went*.

This procedure allowed us to collect incremental cloze values for every word in our 30-min story. Using these values, we characterized the predictability of all the content words (nouns, verbs, and adjectives) in our story. Given the design of the present study, we knew that there would be some words that were highly predictable and others that were not. The question for this analysis, however, was how often the content words were *highly* predictable. Figure 6 shows the distribution of cloze probabilities for all 1,947 content words in our story. The median cloze value was 6.7% (out of 30 total guesses). There were 639 words (roughly 33% of all words) that no participant correctly guessed. Thus, even in this simple edited narrative, most words could not be predicted before they were encountered.

In light of these findings, we conclude that form-based prediction can emerge for highly predictable words despite these words occurring in a broader, more variable environment with generally unpredictable words. Although we cannot ascertain whether our story is more or less predictable than other forms of natural language, we can compare these cloze values with those from the stimuli used in prior psycholinguistic studies. Prior studies that manipulate predictability typically have the following classification: words with cloze values under ~10% are low cloze, words with values up to ~65% are medium cloze, and words with values over 65% are high cloze (Block & Baldwin, 2010; Brothers & Kuperberg, 2021; Kutas & Hillyard, 1984). By this criterion, the content words in our story were mostly low cloze (52.2%) with many medium cloze words (31.2%) and a smaller number of high cloze words (16.6%). The distribution of cloze values in our stimuli also appears to be broadly similar to the values

**Figure 6**

*Distribution of Cloze Probabilities for All Nouns, Verbs, and Adjectives in Our Story*




*Note*. We calculated the cloze probability for all 1,947 content words in our story stimulus. These words are presented by syntactic category: noun (left), verb (middle), or adjective (right). In each panel, the bars represent the total number of words within each cloze bin (ranging from 0% to 100%). ADJ = adjective. See the online article for the color version of this figure.

observed in cloze studies using written passages intended for adults (Lowder et al., 2018; Luke & Christianson, 2016; Smith & Levy, 2013). In sum, these analyses suggest that our story was not unusual in its degree of predictability, and thus, it is unlikely that the effects we observed were driven by a strategy that is specific to the materials that we used.

### *How Slowly Was Our Story Read?*

Most researchers would agree that making top-down predictions during comprehension takes some amount of time (e.g., DeLong et al., 2021; Freunberger & Roehm, 2016; Ito et al., 2016; Pickering & Gambi, 2018; Pickering & Garrod, 2007). To predict the form of an upcoming word, comprehenders must do several things: First, they must perceive the earlier words in an utterance and use them to make inferences at higher conceptual levels about what is likely to come next. Then, after making those inferences, they must transmit information back down to lower levels in order to preactivate (1) the relevant lexical concepts and (2) the form features associated with them (in that order). These steps constitute a feedback loop in which information from the perceived input is propagated to higher levels and then back down again (Dell, 1986; Pickering & Garrod, 2007).

Decades of research on incremental processing have provided insight into how (and when) we should expect to see substantial effects of feedback loops. First, feedback signals should emerge gradually over time, becoming stronger as more time passes (Dell, 1986). Second, in a system with hierarchical, multilayered representations, it should take longer for top-down information to reach lower levels than higher ones (Elman & McClelland, 1988; Indefrey & Levelt, 2004; Pickering & Garrod, 2013; Rumelhart & McClelland, 1982). On this account, we should expect predictions about upcoming forms to emerge later in time than predictions about upcoming meanings or concepts, as form-based prediction requires the same steps as semantic prediction plus the additional step of preactivating phonological and perceptual features. Taken together, these insights have generated skepticism about whether form-based prediction readily occurs during ordinary comprehension (e.g., Freunberger & Roehm, 2016; Indefrey & Levelt, 2004; Ito et al., 2016, 2017a; Pickering & Gambi, 2018; Pickering & Garrod, 2013).

If we assume that comprehenders only begin to predict a given word in an utterance (the target word) after encountering the word immediately before it—and if we assume that this prior word lasts between 200 and 400 ms—then comprehenders only have a few hundred milliseconds to identify the incoming word, generate expectations about the next word, and preactivate its form. There are three reasons for assuming that prediction is limited to the next word: First, if preactivation serves to facilitate the perceptual processing of incoming words, it will only be helpful if that activation is synchronized to the input (i.e., preactivating the form features of a word that is not about to be spoken may do more harm than good); Second, if prediction extends over several words, mechanisms will be required to bind the features of a single word and keep the predictions of distinct words separate; Third, the prior empirical evidence for predictive processing relies on effects observed either at the target word (or immediately before it). For a review of contemporary theories of prediction (with an emphasis on *next*-word prediction), see Ryskin and Nieuwland (2023).

As we noted in the introduction, there is some empirical evidence to support this claim that form-based prediction takes longer than prediction about upcoming meanings or concepts. Specifically, Ito et al. (2016) found reduced N400 responses to unexpected, form-similar words at a slow presentation rate (1.5 words per second) but not at a faster one (two words per second). Given that natural speaking and reading rates are between three and five words per second (Brysbaert, 2019; Tauroza & Allison, 1990), these findings suggest that form-based prediction should rarely occur in everyday contexts.

The stimulus in the present study was an audio recording of a children's story that we intend to also use with child participants. Because child-directed speech is often, but not always, slower than speech directed to adults (e.g., Biersack et al., 2005; Fernald et al., 1989; Ratner, 2013), one might wonder whether our story was so slow that it allowed adults to pursue a predictive strategy that would not ordinarily be available to them. In an earlier section, we calculated the average speech rate for our stories to be 3.25 words per second, which falls within the range of three to five words per second for natural adult-directed speech (Tauroza & Allison, 1990) and well within the range for studies of spoken and written language comprehension in adults (see Dambacher et al., 2012; DeLong et al., 2019, 2021; Ito et al., 2016; Wlotko & Federmeier, 2015).

Moreover, other studies using similar designs and paradigms to Ito et al. (2016) have also found strong evidence for form-based prediction at faster rates (DeLong et al., 2019, 2021; Kim & Lai, 2012; Laszlo & Federmeier, 2009). For example, DeLong et al. (2021) had adults read sentences word-by-word at a rate of four words per second. All of their sentences originally had a highly predictable noun, which they either kept or replaced with an orthographically similar word, a semantically similar word, or an unrelated word, e.g., "The Doberman stood its ground and bared its (*teeth*/*tenth*/*dentist*/*report*) to the mailman." The authors found smaller N400s for the orthographically and semantically similar words relative to the unrelated words—despite presenting their sentences at nearly twice the speed of Ito et al. (2016).

Despite this clear pattern of findings, we are inclined to agree with the skeptics: It seems implausible, given what we know about the mind, that a listener can identify a word as it unfolds, integrate it into a higher level discourse structure, make a prediction for the next word, and preactivate the form of that next word—all in the span of ~300 ms. So, how can we explain the findings from our study, as well as those from the prior studies that report form-based prediction at fast presentation rates?

## Getting a Head Start on Form-Based Prediction

One way to reconcile these findings with our understanding of the temporal properties of feedback loops is to assume that form-based prediction is not typically triggered at the immediately preceding word. Note, *triggering* a prediction is not the same as serving as the basis for that prediction. We still assume that these predictions are based on the entire context up to that point and not on a singular word. Instead, the words that appear earlier in the context may be able to generate or trigger specific lexical expectations that will be fulfilled several words downstream—we will call this *long-distance* prediction. If these long-distance predictions can be made in parallel with the bottom-up processing of each subsequent word, then such a system could allow predictions to emerge gradually during comprehension. In the rest of this section, we review the preliminary evidence for long-distance prediction and then explore the degree to which this phenomenon may account for the divergent findings in prior work. To do this, we investigate whether long-distance prediction might have been possible for the critical words in the present study, as well as two prior studies on form-based prediction.

### *Preliminary Evidence for Long-Distance Prediction*

The clearest evidence for long-distance prediction comes from a recent study exploring form-based prediction in the visual world paradigm. X. Li et al. (2022) asked native Mandarin speakers to look at visual displays while listening to highly constraining sentences like "After school, I put my pencil case and notebooks into my *schoolbag* and get ready to go home." The visual displays always contained four objects, positioned in the four quadrants of the screen. Three of these objects were unrelated distractor objects that shared no semantic or phonological features with the highly predictable noun (e.g., *schoolbag* in the sentence above). The fourth object was either the highly predictable noun (e.g., a schoolbag), a semantic competitor (e.g., an eraser), a phonological competitor (e.g., in Mandarin Chinese, *comb* and *schoolbag* have the same first syllable and tone), or another unrelated distractor (e.g., funnel).

X. Li et al. (2022) found that participants began looking to the semantic and phonological competitors (over the distractors) well before the target word was produced (see also Ito et al., 2018). Specifically, the authors observed increased looks to both competitors starting ~1,400 ms (or two words) before the target word onset. This pattern was interpreted as evidence that form-based predictions were made well in advance of the target word. To determine whether long-distance predictions were possible in their sentence contexts, X. Li et al. (2022) conducted an exploratory cloze task with a different set of Mandarin speakers. These participants heard the original target sentences; however, the authors truncated them ~1,400 ms before the target word. Participants were then asked to complete each of the truncated sentences. Results indicated that participants included the target words in their completions at rates well above chance (e.g., average target cloze probability was 33% with a range of 20%–45%). Thus, it is clear that, under some circumstances, listeners can make predictions about an upcoming word on the basis of information presented much earlier in the sentence—and moreover, these predictions can result in the preactivation of form features well over a second before the predicted word is produced.

### *Evidence for Long-Distance Prediction in the Present Study (and Two Prior Studies)*

This finding raises the possibility that the reduced N400s in form-based prediction studies sometimes result from predictive processing that occurs well before the pretarget word. Specifically, we might expect that the studies showing form-based prediction at faster stimulus presentation rates have stimuli that support long-distance prediction, while the studies showing form-based prediction only at slow presentation rates may not. To explore this, we revisited the stimulus sets used in DeLong et al. (2021) and Ito et al. (2016). DeLong et al. (2021) reported form-based prediction at presentation rates of four words per second, whereas Ito et al. (2016) found that

form-based prediction broke down at rates of two words per second. In addition, we also explored a subset of our target sentences to see if they provided support for long-distance prediction.

To investigate this systematically, we conducted a set of exploratory cloze tasks that presented participants with truncated versions of the high cloze sentences from these three studies (160 high cloze sentences from DeLong et al., 2021; 88 from Ito et al., 2016; and 25 from the present study).[7] For the two reading studies, we presented each target sentence three times, providing additional context each time. For example, we first showed participants a truncated sentence that stopped four words before the target (*The lumberjack chopped* … ) and asked them to complete the sentence. Then, we revealed two words (*The lumberjack chopped the wood* … ) and asked for another completion. Finally, we presented the entire sentence up to the target word (*The lumberjack chopped the wood with his* … ) to determine whether participants could guess the target word *ax*. For the materials from the present study, we implemented a similar procedure: The cartoon would play until reaching four words before the target, and then it would pause. Participants were instructed to then complete the current sentence before receiving more context from the story. This procedure was identical to the one for the two reading studies; thus, we were able to collect comparable incremental cloze values for these various high cloze contexts.

Figure 7 shows the by-item cloze probabilities from each study at the three different distances. In all three studies, participants readily predicted the target words right before they appeared in the input: The present study had an average cloze of 92.7% ($SD = 16.3\%$); the DeLong study had 91.8% ($SD = 12.5\%$); and the Ito study had 88.7% ($SD = 16.4\%$). Long-distance prediction rates, however, varied across the studies. The target words from the present study could often be predicted at distances of four words (*mean cloze* = 39.5%, $SD = 29.9\%$) and two words (*mean cloze* = 70.3%, $SD = 28.0\%$) before they appeared. The two reading studies showed less long-distance prediction; however, the target words in both studies were still predicted at rates above 20% for all distances. The DeLong targets were numerically more predictable across all distances (four words prior, *mean cloze* = 23.3%, $SD = 30.4\%$; two words prior, *mean cloze* = 51.9%, $SD = 35.5\%$) than the Ito targets (four words prior, *mean cloze* = 20.1%, $SD = 27.9\%$; two words prior, *mean cloze* = 48.8%, $SD = 33.1\%$); however, we do not believe that this constitutes a categorical difference in the degree of long-distance prediction in the two reading studies.[8]

In short, these findings provide additional evidence for long-distance prediction, generating a slew of questions about the types of systems that can generate predictions in advance and maintain them as more words are being perceived. We suspect that this kind of long-distance prediction is common in rich naturalistic contexts like our story. Narratives typically follow characters across events and revisit the same topics, making it useful to track (at many levels) the ideas, objects, people, and places that are likely to be mentioned. The cues that allow us to make these predictions may occur a few words, a few sentences, or even a few paragraphs in advance of the target word, allowing for prediction to occur even at quite rapid presentation rates.

## How Does the Present Study Relate to Prior Work on Phonological Mismatch Effects?

In the introduction, we reviewed a handful of studies that investigated the effects of phonological mismatches during spoken language comprehension. In these studies, participants listened to constraining sentences with highly predictable sentence-final words, e.g., "It was a pleasant surprise to find that the car repair bill was only 17 *dollars*." On some trials, these predictable words appeared as expected. On other trials, these words were replaced with semantically incongruous words that either shared the same initial phonemes with the expected word (*dolphins*) or did not (*scholars*). Across these studies, there are three main findings that must be reconciled with the present study, each of which we address below.

First, a handful of studies report early negativities that are argued to be distinct from the N400 for two reasons: (1) they typically emerge and peak earlier than canonical N400s (200–300 ms following the onset of a violation with unexpected initial phonemes) and (2) there is sometimes an apparent temporal discontinuity between an early peak and the peak of the N400 (e.g., Connolly & Phillips, 1994; van den Brink et al., 2001). In the present study, we did not find an early negativity with a peak that was spatially and temporally distinct from our N400 responses. This is not terribly surprising, as auditory ERPs rarely show distinct early effects due to the variability in how spoken words are produced and how they unfold over time (Holcomb & Neville, 1991; Kutas et al., 1987, 2006; Kutas & Van Petten, 1994; Swaab et al., 2012). In a naturalistic stimulus like ours, there is a high level of variability in the onsets of each word, the prosody of each utterance, and the timing between each subsequent word. This variability has the potential to blur together any early effects by increasing noise and interfering with participants' abilities to precisely predict when a word will begin and how it might sound. If we were to use these materials but tightly control the timing of each spoken word (e.g., playing sentences word-by-word or introducing gaps between them), it is possible that we might find distinct early effects (see Kutas & Federmeier, 2011). In our naturalistic *Storytime* paradigm, however, this level of control is not possible, and thus, one limitation of this particular approach is its reduced sensitivity to small, early, and/or short-lived ERP effects.

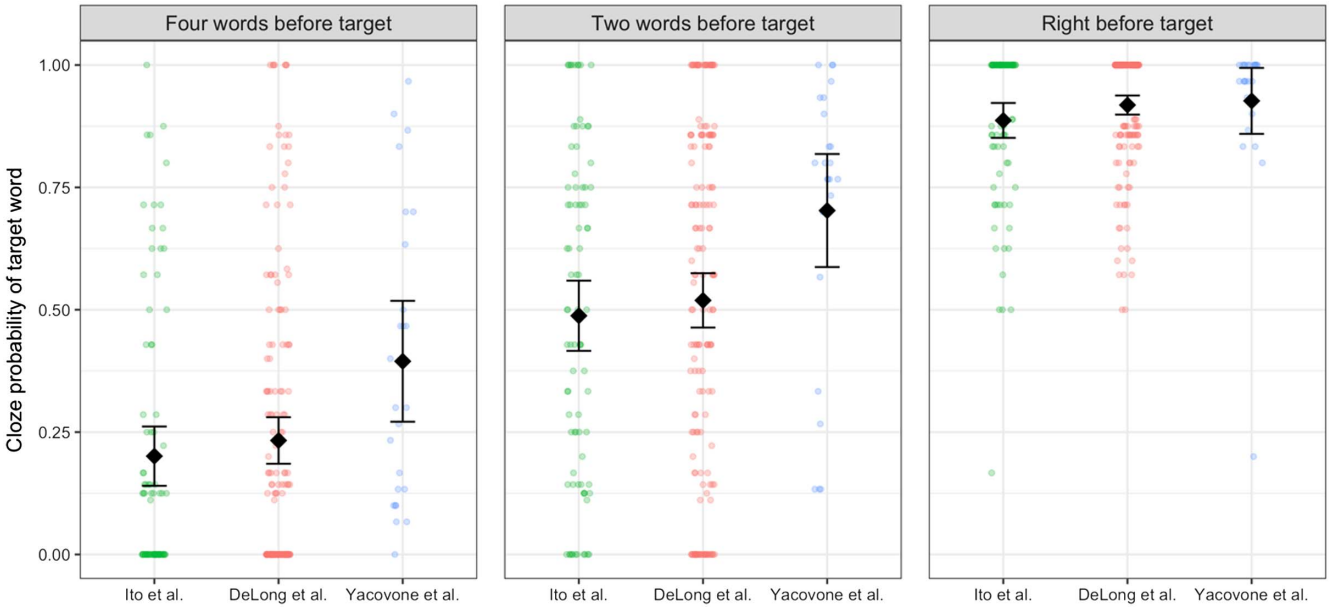
Second, the studies on phonological mismatches during spoken language comprehension typically find a delay in the mismatch effect for violations that share initial phonemes with the expected word (e.g., Liu et al., 2006; Van Petten et al., 1999). For example, Van Petten et al. (1999) found a later N400 effect for

[7] Given the nature of our spoken story, we selected 25 high cloze target words (out of 90) for participants to complete online. For this truncated cloze task, all target sentences needed to have a minimum of five words before the target word. We did not use all viable sentences from our stimulus because participants needed to watch an entire cartoon rather than simply read the sentence, and having them guess 90 target words dramatically extended the duration of the study.

[8] Given these findings, one might wonder why the participants in the Ito study did not show form-based prediction at faster presentation rates, while those in the DeLong study did. We see three possible explanations: First, the small difference in both immediate predictability and long-distance predictability led to more rapid prediction of the critical items in DeLong; Second, because the high predictability sentences formed a much smaller portion of the Ito stimulus set, participants may have engaged in less predictive processing in the Ito study; Third, the discourse constraints allowing for prediction in the Ito study may be more complex and involve higher level conceptual relations that are only calculated when more attentional resources are available (as in our offline cloze task), while the discourse constraints in DeLong might be more associative and less dependent on attention.

**Figure 7**

*Investigating Long-Distance Prediction in Three Studies*




*Note.* The three panels correspond to the point at which the preceding context was truncated (relative to a target word). If the sentence was "*I like my coffee with cream and … sugar*", participants would get "I like my" (four words before target, left panel), "I like my coffee with" (two word before target, middle panel), and then "I like my coffee with cream and" (right before target, right panel). Cloze probabilities were then determined by calculating how often participants provided the target word (sugar) in their sentence completions at each time point. Each point represents a single word (and its cloze probability at varying distances), and there are 160 targets from DeLong et al. (2021), 88 targets from Ito et al. (2016), and 25 targets from Yacovone et al. (the present study). See the online article for the color version of this figure.

unexpected words that shared initial phonemes with the predicted word (e.g., *dolphins* when expecting *dollars*) relative to those that did not (*scholars* when expecting *dollars*). This pattern is consistent with a large body of evidence demonstrating that people incrementally interpret spoken words, restricting their hypotheses about the word's identity as it unfolds (e.g., Allopenna et al., 1998; Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989). In a predictive system with incremental interpretation, we should expect to see early violation effects when the initial phonemes violate our predictions (*scholars*) and later effects when the initial phonemes are consistent with our predictions (*dolphins*).

Thus, it is somewhat surprising that we did not find any latency differences between our form-similar (*ceke*) and less-similar rime (*vake*) violation effects in either cloze condition. In high cloze environments, the N400s for both the *ceke* and *vake* conditions began to diverge from the baseline at ~200 ms (see Figure 2). The only difference between these two effects was in overall amplitude, as the N400 for *ceke* was always smaller than the N400 for *vake* in predictable contexts. In low cloze environments, we did not see any significant differences in amplitude, latency, or scalp distribution for the nonword N400s.

We suspect that these patterns reflect two features of the present study: First, as we mentioned above, the inherent variability when using the *Storytime* paradigm limits our ability to detect small, short-lived effects. Second, because the present study focused on form-based prediction rather than incremental interpretation, both nonword conditions diverged from the target word quite early. Studies of incrementality in spoken language have generally used cohort competitors that have prolonged phonological overlap with the target or expected word (e.g., Allopenna et al., 1998; Liu et al., 2006; Van Petten et al., 1999). For example, Van Petten et al. (1999) had an onset-overlap condition that shared an entire syllable with the expected word (*dollars* vs. *dolphins*). In contrast, their rime-overlap condition immediately mismatched the expected word in initial phonemes (*dollars* vs. *scholars*). Thus, we estimate that Van Petten et al. (1999) had a period of roughly 300 ms (out of a total word duration of ~585 ms) in which the rime-overlap violation was detectable but the onset-overlap violation was not.

In contrast, the present study did not have long periods of time in which listeners could detect the rime nonwords but not the form-similar nonwords. The form-similar violations only shared an initial consonant (or consonant cluster) with the target word and then diverged at the initial vowel (e.g., *cake* vs. *ceke*). The rime violations had the opposite pattern, diverging from the target word at the initial consonant(s) but sharing an initial vowel (e.g., *cake* vs. *vake*). As a result, the form-similar violations could presumably be detected shortly after the release for stop consonants (or even on the basis of co-articulatory cues for nonstop consonants). In other words, the time needed to disambiguate *cake* from *ceke* and *cake* from *vake* probably differs by tens of milliseconds rather than hundreds like in prior studies. Thus, any delay in violation effects for *ceke* would have been short lived, making it difficult to detect using the present paradigm.

The final issue to explore is why some spoken language studies demonstrate reduced N400s for onset-overlap violations (e.g., van den Brink et al., 2001), while others do not (Van Petten et al., 1999). In contrast to the present study, Van Petten et al. (1999) did not find reduced N400s to onset-overlap violations (relative to rime-overlap violations) after controlling for differences in disambiguation points. This prior study and our study used nearly identical violation designs but found different patterns of results—thus, we must attempt to reconcile these two studies to better understand the conditions in which form-based prediction occurs.

One clear difference between these two studies is the use of nonword versus real-word violations (in addition to the clear differences in disambiguation points described above). When a listener encounters a nonword like *ceke* in an environment that is highly constraining for the word *cake*, they may be apt to simply interpret this input as being consistent with *cake*. In contrast, if a listener encounters a real-word violation like *dolphins* in a context that predicts *dollars*, it may be more difficult to recover the intended meaning of the utterance and re-cast *dolphins* as meaning *dollars*. Thus, if the size of the N400 reflects the amount of additional processing needed to override a prior prediction, activate the newly perceived lexicosemantic features, and integrate the unexpected word into the context, then it makes sense that *dolphins* and *scholars* evoke similarly sized N400s. However, if the violation neither brings new lexicosemantic features nor strongly disconfirms a prior prediction (e.g., *ceke*), then it should be processed more similarly to the expected word than something unexpected. This hypothesis predicts that nonword and real-word violations should be processed differently in spoken language comprehension.

Preliminary support for this hypothesis comes from a study of spoken language comprehension in Mandarin Chinese. In this study, Liu et al. (2006) had two experiments: the first used real-word violations similar to Van Petten et al. (1999), whereas the second used nonword violations similar to the present study.

In Experiment 1, native Mandarin speakers listened to sentences with highly predictable endings, e.g., "The sound in the radio became weaker and weaker. It seems that I must buy several new sets of *batteries*." Similar to prior work, the authors manipulated the last word to be the expected word (*batteries*, in Mandarin ***/dian4/-/chi2/***) or one of three real-word violations: an onset-overlap violation (*electric stove*, ***/dian4/-/lu2/***), a rime-overlap violation (*water pool*, */shui3/-/chi2/*), or a no-overlap violation (*illness*, */bing4/-/tai4/*). They found increased N400 amplitudes for all violations relative to the expected word—and critically, the timing of these effects differed from one another, but the overall amplitude did not. Specifically, the N400 effect for the onset-overlap (***/dian4/-/chi2/*** vs. ***/dian4/-/lu2/***) emerged later than the N400 effects for the other violations.

In Experiment 2, another set of native Mandarin speakers listened to both predictable and unpredictable sentences. The authors manipulated the last word in these sentences to be one of four conditions: the *original word* from the sentence; a *minimal-onset-mismatch* (a nonword with an onset that mismatches the original word in one or two features); a *maximal-onset-mismatch* (a nonword with an onset that mismatches the original by two or more features); and a *first-syllable-mismatch* (a nonword with a completely different first syllable than the original word, i.e., the rime condition from prior studies). Similar to the present study, Liu et al. found N400 effects to all nonword violations—however, in predictable contexts, the size of the N400 effect depended on the degree of form similarity to the expected word: the N400s for *original word* < *minimal-onset-mismatch* < *maximal-onset-mismatch* < *first-syllable-mismatch*. In the unpredictable contexts, all three nonword violations produced similarly sized N400 effects. Taken together, these findings tentatively support the hypothesis that real-word and nonword violations produce different ERP effects in spoken language comprehension.

## How Does the Present Study Advance Our Understanding of Late Posterior Positivities?

In the introduction, we mentioned that posterior P600s often accompany N400 effects in studies of form-based prediction. We replicated this finding in the present study; however, our results slightly diverge from the prior literature in two ways:

First, we observed P600 effects for *both* form-similar (*ceke*) and less-similar (*vake*) violations, and these effects were similar in magnitude. Prior work has shown that the degree of form similarity between a violation and an expected target word influences the size of the P600 (e.g., Ito et al., 2016; Laszlo & Federmeier, 2009; Ryskin et al., 2021; Vissers et al., 2006). So, why might our two violations elicit similarly sized P600s? While *cake* may seem more similar to *ceke* than *vake* when processed incrementally, both violations only differ from the target word by one or two phonemes. After all, we decided to use rime violations in order to allow listeners to recover the intended meaning of utterances with violations in them. Thus, a simple explanation for why the P600s may be similar in size for both *ceke* and *vake* could be that the P600 reflects a process that occurs *after* bottom-up processing is complete (for similar discussion, see Ito et al., 2016). If the processes indexed by the P600 are *reactive* rather than predictive, both *ceke* and *vake* can be construed as slight deviations from a more congruent continuation (*cake*). On this account, we might expect that comprehenders will face similar difficulties when attempting to incorporate these two nonwords into their higher level interpretations and when reprocessing these violations in order to assess the nature of the anomalous input.

Second, we observed P600 effects for all violations regardless of the predictability of the target word. Prior work has demonstrated that P600s are more robust in high constraint contexts (Gunter et al., 2000; Kuperberg et al., 2020; van de Meerendonk et al., 2010; Van De Meerendonk et al., 2009). This finding has led some researchers to argue that the P600s in the form-based prediction literature reflect comprehenders' interpretation of the violations as misspellings of the predicted word (which also explains why P600s are not readily observed in low constraint contexts; see Vissers et al., 2006). So, why did we observe robust P600 effects in our less predictable conditions? We see two possible explanations for this pattern.

The first possibility is that the P600 simply reflects the initial failure to incorporate the bottom-up input into one's high-level interpretation of the context, as well as the subsequent disruption to ongoing comprehension caused by reprocessing the anomalous input (Brothers et al., 2020, 2022; Kuperberg, 2007; Kuperberg et al., 2020). On this account, comprehenders do not need to have strong top-down expectations about what is coming next in the sentence—however, they must be actively engaged in deep comprehension to experience a disruption from the violation (Brothers et al., 2020). Typically, low constraint contexts do not provide rich details about an unfolding event; thus, deep comprehension may be difficult

to achieve in these kinds of sentences. In the present study, however, the unpredictable sentences are embedded within a larger discourse. So, although a sentence may be unpredictable at the local level, it is still contributing to the understanding of the broader context. On this account, we would expect P600s to violate these unpredictable sentences because comprehenders are deeply processing the linguistic material, and the violations are disrupting the ongoing construction of the narrative.

The second possibility is that there are some sentences in the low cloze group that are actually highly constraining for a different word than the one in the story. For example, if someone said, "For my birthday, I am going to bake a large *pie*," you would not consider the sentence to be anomalous despite it violating your expectations. In these scenarios, the context is generating strong constraints for a particular continuation (e.g., *cake*); however, the speaker never produces the predicted word. Thus, if the P600 is primarily sensitive to anomalous (or highly implausible) continuations in high constraint contexts, then we should expect to see robust P600s in our low cloze environments when they strongly constrain for an unobserved alternative word.

To investigate this, we characterized the constraint of all sentences containing target words (regardless of the cloze probability of the word from the story). For this analysis, we focused on low cloze target words (however, Figure 8 presents these exploratory findings alongside the high cloze condition). All of our low cloze items had target words with cloze values less than 50% (out of 30 total responses). To measure the constraint for each item, we calculated the cloze probability of the most frequent response produced by participants in our cloze task. We then grouped these items using a median split approach: Sentences with constraint values greater than 40% were classified as high constraint, and those with values less than 40% were classified as low constraint.

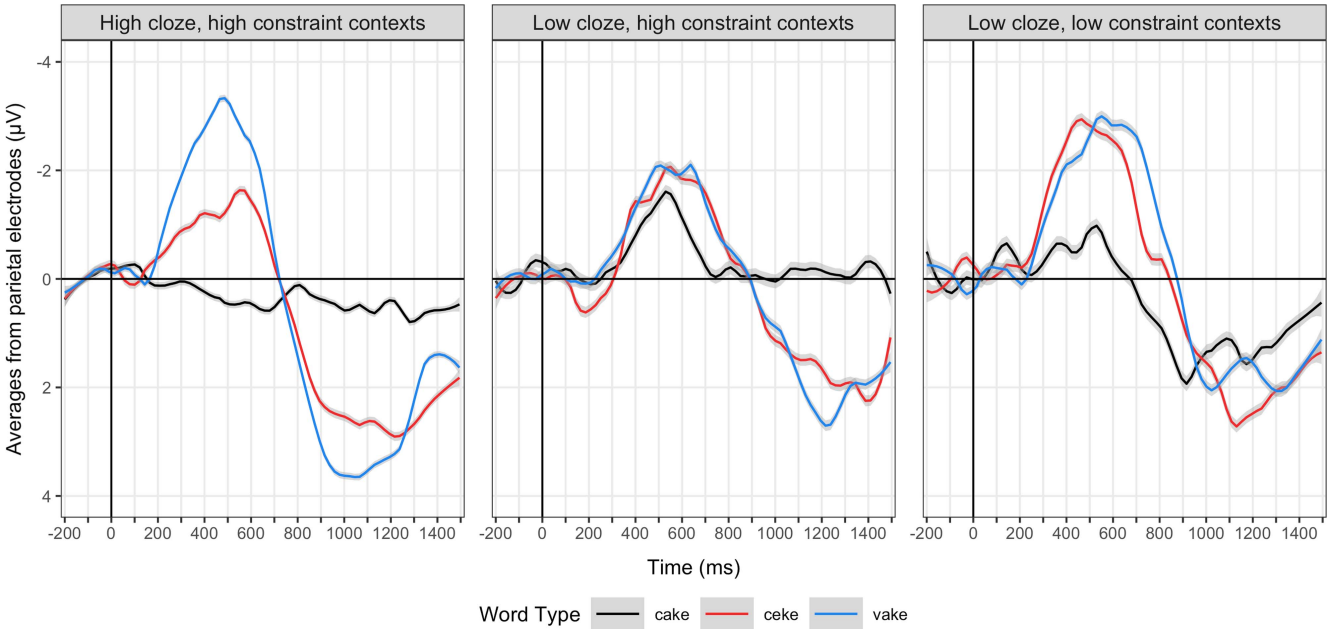
In Figure 8, we visualize the grand average waveforms for three groups of words: high cloze target words in high constraint contexts, e.g., "I like my coffee with cream and *sugar*" (left panel); low cloze target words in high constraint contexts, e.g., "For my birthday, I wanted to bake a large *pie*" (middle panel); and finally, low cloze target words in low constraint contexts, e.g., "They are only found in a certain *river*" (right panel). These waveforms revealed robust P600 effects for all violation conditions in high constraint contexts (regardless of the cloze probability of the target word). These observations tentatively support the claim that the P600 effects in our low cloze conditions (see Figure 2, bottom panel) were largely driven by sentences with high cloze competitors.

## What Are the Open Questions and What Should We (Collectively) Do Next?

The present study demonstrates that form-based prediction is a widespread phenomenon in language comprehension, occurring in both tightly controlled experiments and in more variable naturalistic contexts. If prediction is as ubiquitous as our data suggest, a number of unanswered questions take on new importance: How is prediction carried out in linguistically diverse populations such as signers and bilinguals? How is prediction affected by cognitive variability (e.g., autism, schizophrenia, or attention-deficit/hyperactivity disorder)? How is it affected by aging? What is the relationship between variation in predictive abilities and variation in other measures of

**Figure 8**

*Visualization of the ERP Effects Across Sentence Constraint and Target Predictability*




*Note*. The grand average waveforms (μV) from parietal electrodes (*Pz, P3, P4*) are plotted in three groups, depending on the target word's cloze probability (high or low cloze) and the overall constraint of the context (high or low constraint). The black lines represent the baseline condition (*cake*), and the red and blue lines represent the form-similar (*ceke*) and rime (*vake*) conditions, respectively. These waveforms were produced in R and then smoothed using local regression (loess) smoothing techniques. ERP = event-related potential. See the online article for the color version of this figure.

linguistic ability? How do we represent a form-based prediction while processing the form of the current (perceived) word? This latter point seems like a particularly tricky problem if we engage in long-distance predictions, as this could well involve the simultaneous prediction of a number of different lexical items.

One unanswered question, however, seems particularly urgent: How and when does form-based prediction develop? Most of the prior work on form-based prediction has recruited highly competent language users (e.g., adults with high levels of education and literacy). As we noted above, form-based prediction requires the rapid coordination of several processing steps. Comprehenders must leverage contextual information to make inferences at the highest conceptual levels and then send information back down to preactivate lower level representations—all before the critical word is produced and perceived. Thus, effective prediction should require considerable knowledge about one's language (and the world, more generally), as well as a rapid and efficient processing system. Given that young children are both less knowledgeable and slower at basic cognitive tasks (Kail, 1991; R. V. Kail & Ferrer, 2007; Kail & Salthouse, 1994), one might expect that they would be less apt to make form-based predictions. This deficit could resolve gradually as processing speed increases and children fill in the gaps in their knowledge. Or there could be a sudden shift in the system as children adopt different processing strategies—either as a side effect of their effectiveness (e.g., ignoring predictions until they have been found to be accurate) or based on their experiences with literacy (Huettig & Pickering, 2019; Mani & Huettig, 2012).

To date, most of the research on predictive processing in children has focused on semantic prediction using the visual world paradigm. For example, Mani and Huettig (2012) presented German-speaking 2-year-olds with visual displays that had two objects (e.g., a cake and a bird). On each trial, the children would hear a sentence that either contained a neutral verb or a highly constraining verb. For example, "The boy eats (sees) the big cake" (German translation, "Der Junge ißt (sieht) den großen Kuchen"). The authors found that the children made predictive eye movements to the critical object (the cake)—but only after hearing the highly constraining verb (eats). These and other similar findings suggest that young children can use contextual cues to generate predictions (see Borovsky et al., 2012; Kidd et al., 2011; Lew-Williams & Fernald, 2007).

These findings, however, have two general limitations: First, it is unclear from the dependent measure in these studies whether children are making predictions about what is going to be said next or simply making inferences about the event under discussion. Prior work has shown that adults and children will look toward objects that are contextually relevant—even when they presumably know that this object is unlikely to be explicitly mentioned. For example, when given *wh*-questions like "What did the man eat?" people will look longer at edible objects in the scene (Atkinson et al., 2018; Golinkoff et al., 2013; Goodwin et al., 2012; Jyotishi et al., 2017; Seidl et al., 2003; Sussman & Sedivy, 2003; Yuan et al., 2011). Second, even if predictive eye movements are generated by a linguistic prediction, it is unclear whether the prediction is at the level of meaning and/or form, i.e., are 2-year-old children activating the concept of *cake*, the lexical item *cake*, or the phonological form of that lexical item?

To the best of our knowledge, there is only one study that directly explores form-based prediction in young children. Gambi et al. (2018) conducted a visual world eye-tracking study with English-speaking adults and children (2–5 years old). On the critical trials, the visual display consisted of two objects, each one positioned on a different side of the screen. One object would typically be labeled by a word beginning with a vowel (e.g., an ice cream cone). The other object would typically be labeled by a word beginning with a consonant (e.g., a soccer ball). In the test sentence, the two pictures were labeled with an indefinite determiner that provided predictive information about the identity of the upcoming noun (e.g., "Can you see *an* … ice cream?"). Participants' performance on these form-based prediction trials was contrasted with their performance on trials in which different numbers of objects appeared on each side of the display and a number word was used in the test sentence (e.g., "Can you see *two* … ice creams?"). To provide more time for making predictions, the authors inserted a pause after the determiner such that the target word was produced roughly 1,200 ms later.

In the number trials, results indicated that all age groups were able to shift their gaze to the correct referent after hearing the number word. Gambi et al. (2018) interpreted this finding as semantic prediction—although, it could also be interpreted as a product of incremental semantic analysis (i.e., participants shift to looking at sets of objects after hearing *two*, just as they might shift to look at green objects after hearing *green*). In the form-based prediction trials, the youngest children failed to shift their gaze on the basis of the indefinite article, suggesting that they were unable to use the phonological cues from the determiner to predict the phonological form of the upcoming noun and then correctly infer the referent. Three- to 5-year-old children showed fragile effects of form-based prediction across their analyses, which led the authors to conclude that form-based prediction was only reliably observed in adults.

Based on this finding, one might conclude that, even in the most supportive of contexts, form-based prediction is absent until at least 5 years of age (see Pickering & Gambi, 2018). We suspect, however, that this conclusion is premature, and we see two reasons why additional research is needed: First, the phenomenon at the heart of Gambi et al. (2018) appears to be a rather weak one that is atypical of form-based prediction more broadly. In their study, words were only predictable because of arbitrary phonological rules that allowed participants to predict the onset of an upcoming word and then infer the correct referent. In contrast, when a word was predictable in the present study, it was because the content of the discourse constrained the possibilities of which words could or should come next, which in turn constrained expectations for possible word forms. Thus, form-based prediction in our study was based on a top-down flow of information, whereas prediction in the Gambi study relied on the form of one word constraining the form of the next.

We suspect that this latter pathway is rarely useful in the wild: Phonological constraints are often highly local cues (e.g., the a/an distinction requires attention to immediately adjacent phonemes). Thus, there is little time to use these constraints to generate predictions in real time. Furthermore, predictions that are based solely on phonological information would generally be quite coarse, as there are typically just two or three alternative forms of the predictive cue (resulting in thousands of possible lexical candidates). Gambi et al. (2018) carefully designed their study to overcome these real-world impediments by reducing the discourse context to just two objects and using artificially long prosodic breaks to afford more time for prediction. It is remarkable that adults in this study were able to flexibly adapt to these circumstances, but it is not surprising that children were less flexible. While the a/an constraint is probably the

best known test case for form-based prediction, the findings have been variable, and the effects, if they exist, appear to be quite weak (for additional context, see DeLong et al., 2005, 2017; Ito et al., 2017a, 2017b). We suspect that these inconsistent findings have less to do with the fragility of form-based prediction and more to do with the minimal incremental value of this particular predictive cue.

Second, the wider literature on children's language processing strongly suggests that, similar to adults, children have robust predictive abilities. While none of these studies provide direct evidence for form-based prediction, each of the relevant findings suggests that lexical prediction is common in children, and that it is not radically different from lexical prediction in adults. Thus, if lexical prediction in adulthood involves form-based prediction, then these findings suggest that children engage in form-based prediction as well.

For example, like adults, children show robust N400 effects during comprehension, which suggests that some degree of lexicosemantic preactivation emerges in adolescence (e.g., Friedrich & Friederici, 2006; Henderson et al., 2011; Juottonen et al., 1996). Prior work using the *Storytime* paradigm has shown that N400 responses in adults and children (5–10 years old) are best predicted by the cloze probability of a word (Levari & Snedeker, 2024). This finding suggests that both age groups are sensitive to the predictability of a word given its context. Finally, developmental researchers have long used versions of the cloze task to assess children's comprehension and morphological productivity (Berko, 1958; Brown & Berko, 1960; Carroll, 1971; Shanahan et al., 1982; Skarakis-Doyle & Dempsey, 2008). In these studies, children must correctly predict a word, including its precise phonological form, in order to accurately complete the sentence.

These studies, however, clearly stop short of demonstrating that children often (or ever) use top-down contextual information to predict upcoming words in real time. For this reason, we are currently conducting a study parallel to the present experiment with young children (5–6 years old) to see if they also show reduced N400 effects to form-similar nonwords in highly predictable contexts. The findings of this study will place hard constraints on our theory of how this central skill develops.

## Conclusion

The present study demonstrates that form-based prediction is a widespread phenomenon in language comprehension, occurring in both tightly controlled experiments, as well as in more variable naturalistic contexts. To do this, we relied on a novel naturalistic listening task in which participants simply listened to a narrative with experimental manipulations spliced into it. Our findings suggest that adults were actively predicting the phonological form of upcoming words, as evidenced by reduced N400 responses to form-similar violations in highly predictable contexts. In addition, we observed robust late posterior positivities to all violations in our task, suggesting that adults engage in deep comprehension while listening to naturally produced stories. The success of this paradigm in eliciting robust prediction opens the door for testing predictive abilities in a wide range of populations and age groups. With these findings, we conclude that form-based prediction is a common aspect of comprehension in the wild, and future research must begin to characterize how this phenomenon emerges across development and across a wider range of contexts.

## Constraints of Generality

In the present study, we specifically recruited native English-speaking adults—and thus, our findings (when taken in isolation) can only characterize the phenomenon of form-based prediction in this population. We selected native English-speaking adults for a few reasons: First, English is the target language that has been predominantly used in many prior studies on form-based prediction in the literature. For this reason, we had clear predictions about the magnitude and directionality of our effects. Second, English is the dominant language used by the community in the Greater Boston area, as well as by the study team. We acknowledge that English-speaking adults are often overstudied in psycholinguistics research, and thus, in future work, we aim to characterize similar phenomena in different age groups (e.g., toddlers, school-aged children, teenagers), as well as in different language families and linguistic modalities (e.g., American Sign Language).

## References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Atkinson, E., Wagers, M. W., Lidz, J., Phillips, C., & Omaki, A. (2018). Developing incrementality in filler-gap dependency processing. *Cognition*, *179*, 132–149. https://doi.org/10.1016/j.cognition.2018.05.022

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

Barr, D. (2021). *Learning statistical models through simulation in R: An interactive textbook*. (Version 1.0.0) [Computer software]. https://psyteachr.github.io/stat-models-v1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv. https://arxiv.org/abs/1406.5823

Beres, A. M. (2017). Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research. *Applied Psychophysiology and Biofeedback*, *42*(4), 247–255. https://doi.org/10.1007/s10484-017-9371-3

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Biersack, S., Kempe, V., & Knapton, L. (2005). Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. *Interspeech*, *2005*, 2401–2404. https://doi.org/10.21437/Interspeech.2005-46

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670. https://doi.org/10.3758/BRM.42.3.665

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. https://doi.org/10.1016/j.jecp.2012.01.005

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken

language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 607–624. https://doi.org/10.3758/s13415-015-0340-0

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, *14*(1), Article e0207741. https://doi.org/10.1371/journal.pone.0207741

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, Article 104174. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149. https://doi.org/10.1016/j.cognition.2014.10.017

Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, *1*(1), 135–160. https://doi.org/10.1162/nol_a_00006

Brothers, T., Zeitlin, M., Perrachione, A. C., Choi, C., & Kuperberg, G. (2022). Domain-general conflict monitoring predicts neural and behavioral indices of linguistic error processing during reading comprehension. *Journal of Experimental Psychology: General*, *151*(7), 1502–1519. https://doi.org/10.1037/xge0001130

Brown, R., & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, *31*(1), 1–14. https://doi.org/10.2307/1126377

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, Article 104047. https://doi.org/10.1016/j.jml.2019.104047

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments & Computers*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Carroll, J. B. (1971). *Defining language comprehension*. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge* (pp. 1–29). John Wiley and Sons.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

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 Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

DeLong, K. A., Chan, W. H., & Kutas, M. (2019). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*, *56*(4), Article e13312. https://doi.org/10.1111/psyp.13312

DeLong, K. A., Chan, W. H., & Kutas, M. (2021). Testing limits: ERP evidence for word form preactivation during speeded sentence reading. *Psychophysiology*, *58*(2), Article e13720. https://doi.org/10.1111/psyp.13720

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*(12), 631–645. https://doi.org/10.1111/lnc3.12093

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

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, *32*(8), 966–973. https://doi.org/10.1080/23273798.2017.1279339

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

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*(2), 143–165. https://doi.org/10.1016/0749-596X(88)90071-X

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

Federmeier, K. D. (2022). Connecting and considering: Electrophysiology provides insights into comprehension. *Psychophysiology*, *59*(1), Article e13940. https://doi.org/10.1111/psyp.13940

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Freunberger, D., & Roehm, D. (2016). Semantic prediction in language comprehension: Evidence from brain potentials. *Language, Cognition and Neuroscience*, *31*(9), 1193–1205. https://doi.org/10.1080/23273798.2016.1205202

Freunberger, D., & Roehm, D. (2017). The costs of being certain: Brain potential evidence for linguistic preactivation in sentence processing. *Psychophysiology*, *54*(6), 824–832. https://doi.org/10.1111/psyp.12848

Friedrich, M., & Friederici, A. D. (2006). Early N400 development and later language acquisition. *Psychophysiology*, *43*(1), 1–12. https://doi.org/10.1111/j.1469-8986.2006.00381.x

Gambi, C., Gorrie, F., Pickering, M. J., & Rabagliati, H. (2018). The development of linguistic prediction: Predictions of sound and meaning in 2- to 5-year-olds. *Journal of Experimental Child Psychology*, *173*, 351–370. https://doi.org/10.1016/j.jecp.2018.04.012

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–339. https://doi.org/10.1177/1745691613484936

Goodwin, A., Fein, D., & Naigles, L. R. (2012). Comprehension of wh-questions precedes their production in typical development and autism spectrum disorders. *Autism Research*, *5*(2), 109–123. https://doi.org/10.1002/aur.1220

Gunter, T., Friederici, A., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience*, *12*(4), 556–568. https://doi.org/10.1162/089892900562336

Hagoort, P., & Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research*, *28*(6), 715–728. https://doi.org/10.1023/A:1023277213129

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis. Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194–205. https://doi.org/10.1162/089892999563328

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), Article e2201968119. https://doi.org/10.1073/pnas.2201968119

Henderson, L. M., Baseler, H. A., Clarke, P. J., Watson, S., & Snowling, M. J. (2011). The N400 effect in children: Relationships with comprehension, vocabulary and decoding. *Brain and Language*, *117*(2), 88–99. https://doi.org/10.1016/j.bandl.2010.12.003

Holcomb, P. J., & Neville, H. J. (1991). Natural speech processing: An analysis using event-related brain potentials. *Psychobiology*, *19*(4), 286–300. https://doi.org/10.3758/BF03332082

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition*, *224*, Article 105050. https://doi.org/10.1016/j.cognition.2022.105050

Huettig, F., & Pickering, M. J. (2019). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Sciences*, *23*(6), 464–475. https://doi.org/10.1016/j.tics.2019.03.008

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Ito, A., Corley, M., Pickering, M., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171. https://doi.org/10.1016/j.jml.2015.10.007

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017a). Why the A/AN prediction effect may be hard to replicate: A rebuttal to DeLong, Urbach & Kutas (2017). *Language, Cognition and Neuroscience*, *32*(8), 974–983. https://doi.org/10.1080/23273798.2017.1323112

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017b). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954–965. https://doi.org/10.1080/23273798.2016.1242761

Ito, A., Pickering, M. J., & Corley, M. (2018). Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, *98*, 1–11. https://doi.org/10.1016/j.jml.2017.09.002

Juottonen, K., Revonsuo, A., & Lang, H. (1996). Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. *Cognitive Brain Research*, *4*(2), 99–107. https://doi.org/10.1016/0926-6410(96)00022-5

Jyotishi, M., Fein, D., & Naigles, L. (2017). Investigating the grammatical and pragmatic origins of wh-questions in children with autism spectrum disorders. *Frontiers in Psychology*, *8*, Article 319. https://doi.org/10.3389/fpsyg.2017.00319

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490–501. https://doi.org/10.1037/0033-2909.109.3.490

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*(2–3), 199–225. https://doi.org/10.1016/0001-6918(94)90003-5

Kail, R. V., & Ferrer, E. (2007). Processing speed in childhood and adolescence: Longitudinal models for examining developmental change. *Child Development*, *78*(6), 1760–1770. https://doi.org/10.1111/j.1467-8624.2007.01088.x

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kappenman, E. S., & Luck, S. J. (2011). ERP components: The ups and downs of brainwave recordings. In E. S. Kappenman & S. J. Luck (Eds.), *The Oxford handbook of event-related potential components* (pp. 4–30). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.013.0014

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The language experience and proficiency questionnaire (LEAP-Q): Ten years later. *Bilingualism: Language and Cognition*, *23*(5), 945–950. https://doi.org/10.1017/S1366728919000038

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934. https://doi.org/10.1111/j.1467-7687.2011.01049.x

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

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

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

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*(4), 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195395518.003.0065

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

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*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., Federmeier, K. D., & Urbach, T. P. (2014). The "negatives" and "positives" of prediction in language. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The cognitive neurosciences* (5th ed., pp. 649–656). MIT Press. https://doi.org/10.7551/mitpress/9504.003.0071

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

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (2019). Word expectancy and event-related brain potentials during sentence processing. *Preparatory states and processes*. Psychology Press. https://doi.org/10.4324/9781315792385-11

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology. Supplement*, *39*, 325–330. https://pubmed.ncbi.nlm.nih.gov/3477442/

Kutas, M., & Van Petten, C. K. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). Academic Press.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 659–724). Elsevier. https://doi.org/10.1016/B978-012369374-7/50018-3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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

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

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lenth, R. (2024). *Emmeans: Estimated marginal means, aka least-squares means* (R Package Version 1.10.5). https://rvlenth.github.io/emmeans/

Levari, T., & Snedeker, J. (2024). Understanding words in context: A naturalistic EEG study of children's lexical processing. *Journal of Memory and Language*, *137*, Article 104512. https://doi.org/10.1016/j.jml.2024.104512

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–198. https://doi.org/10.1111/j.1467-9280.2007.01871.x

Lewendon, J., Mortimore, L., & Egan, C. (2020). The phonological mapping (mismatch) negativity: History, inconsistency, and future direction. *Frontiers in Psychology*, *11*, Article 1967. https://doi.org/10.3389/fpsyg.2020.01967

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W., Spreng, R., Brennan, J., Yang, Y., Pallier, C., & Hale, J. (2021). *Le Petit Prince: A multilingual fMRI corpus using ecological stimuli*. bioRxiv. https://doi.org/10.1101/2021.10.02.462875

Li, X., Li, X., & Qu, Q. (2022). Predicting phonology in language comprehension: Evidence from the visual world eye-tracking task in Mandarin Chinese. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(5), 531–547. https://doi.org/10.1037/xhp0000999

Liu, Y., Shu, H., & Wei, J. (2006). Spoken word recognition in context: Evidence from Chinese ERP analyses. *Brain and Language*, *96*(1), 37–48. https://doi.org/10.1016/j.bandl.2005.08.007

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

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*(S4), 1166–1183. https://doi.org/10.1111/cogs.12597

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

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

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

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585. https://doi.org/10.1037/0096-1523.15.3.576

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, *31*(4), 536–548. https://doi.org/10.1080/23273798.2015.1117117

Morgan-Short, K., & Tanner, D. (2013). Event-related potentials (ERPs). In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 127–152). Routledge. https://doi.org/10.4324/9780203123430

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

Nour Eddine, S., Brothers, T., & Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. *Psychology of Learning and Motivation*, *76*, 123–206. https://doi.org/10.1016/bs.plm.2022.03.005

Ochshorn, R., & Hawkins, M. (2016). *Gentle, A robust yet lenient forced aligner built onKaldi* [Computer Software]. https://github.com/lowerquality/gentle/

Osterhout, L., Allen, M. D., McLaughlin, J., & Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Memory & Cognition*, *30*(8), 1304–1312. https://doi.org/10.3758/BF03213412

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786–803. https://doi.org/10.1037/0278-7393.20.4.786

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology*, *52*(11), 1456–1469. https://doi.org/10.1111/psyp.12515

Payne, B. R., Ng, S., Shantz, K., & Federmeier, K. D. (2020). Chapter Four—Event-related brain potentials in multilingual language processing: The N's

and P's. In K. D. Federmeier & H.-W. Huang (Eds.), *Psychology of learning and motivation* (Vol. 72, pp. 75–118). Academic Press. https://doi.org/10.1016/bs.plm.2020.03.003

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110. https://doi.org/10.1016/j.tics.2006.12.002

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

Poulton, V. R., & Nieuwland, M. S. (2022). Can you hear what's coming? Failure to replicate ERP evidence for phonological prediction. *Neurobiology of Language*, *3*(4), 556–574. https://doi.org/10.1162/nol_a_00078

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ratner, N. B. (2013). Why talk with children matters: Clinical implications of infant- and child-directed speech research. *Seminars in Speech and Language*, *34*(4), 203–214. https://doi.org/10.1055/s-0033-1353449

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509. https://doi.org/10.3758/BF03214555

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Memory & Cognition*, *18*(4), 367–379. https://doi.org/10.3758/BF03197126

Rumelhart, D., & McClelland, J. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94. https://doi.org/10.1037/0033-295X.89.1.60

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052. https://doi.org/10.1016/j.tics.2023.08.003

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, Article 107855. https://doi.org/10.1016/j.neuropsychologia.2021.107855

Seidl, A., Hollich, G., & Jusczyk, P. W. (2003). Early understanding of subject and object wh-questions. *Infancy*, *4*(3), 423–436. https://doi.org/10.1207/S15327078IN0403_06

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*(2), 229–255. https://doi.org/10.2307/747485

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2023). *Afex: Analysis of factorial experiments* (R package Version 1.2-1) [Computer software]. https://CRAN.R-project.org/package=afex

Skarakis-Doyle, E., & Dempsey, L. (2008). Assessing story comprehension in preschool children. *Topics in Language Disorders*, *28*(2), 131–148. https://doi.org/10.1097/01.TLD.0000318934.54548.7f

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sussman, R. S., & Sedivy, J. (2003). The time-course of processing syntactic dependencies: Evidence from eye movements. *Language and Cognitive Processes*, *18*(2), 143–163. https://doi.org/10.1080/01690960143000498

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related ERP components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 397–439). Oxford University Press.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, *11*(1), 90–105. https://doi.org/10.1093/applin/11.1.90

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

The MathWorks Inc. (2022). *MATLAB version: 9.13.0* (R2022b). https://www.mathworks.com

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). *Irvine phonotactic online dictionary, Version 2.0*. [Data file]. https://www.iphod.com

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

van de Meerendonk, N., Kolk, H. H. J., Vissers, C. Th. W. M., & Chwilla, D. J. (2010). Monitoring in language perception: Mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, *22*(1), 67–82. https://doi.org/10.1162/jocn.2008.21170

Van De Meerendonk, N., Kolk, H. H. J., Chwilla, D. J., & Vissers, C. Th. W. M. (2009). Monitoring in language perception. *Language and Linguistics Compass*, *3*(5), 1211–1224. https://doi.org/10.1111/j.1749-818X.2009.00163.x

van den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*(7), 967–985. https://doi.org/10.1162/089892901753165872

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255. https://doi.org/10.1016/j.cogbrainres.2004.09.002

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394–417. https://doi.org/10.1037/0278-7393.25.2.394

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., & Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In G. B. Simpson (Ed.), *Understanding word and sentence* (pp. 129–174). North-Holland. https://doi.org/10.1016/S0166-4115(08)61532-0

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

Vissers, C. T. W. M., Chwilla, D. J., & Kolk, H. H. J. (2006). Monitoring in language perception: The effect of misspellings of words in highly constrained sentences. *Brain Research*, *1106*(1), 150–163. https://doi.org/10.1016/j.brainres.2006.05.012

Wang, L., Wlotko, E., Alexander, E., Schoot, L., Kim, M., Warnke, L., & Kuperberg, G. R. (2020). Neural evidence for the prediction of animacy features during language comprehension: Evidence from MEG and EEG representational similarity analysis. *The Journal of Neuroscience*, *40*(16), 3278–3291. https://doi.org/10.1523/JNEUROSCI.1733-19.2020

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *68*, 20–32. https://doi.org/10.1016/j.cortex.2015.03.014

Yacovone, A., Moya, E., & Snedeker, J. (2021). Unexpected words or unexpected languages? Two ERP effects of code-switching in naturalistic discourse. *Cognition*, *215*, Article 104814. https://doi.org/10.1016/j.cognition.2021.104814

Yacovone, A., Waite, B., Levari, T., & Snedeker, J. (2024, April 25). *Let them eat ceke: An EEG study of form-based prediction in rich naturalistic contexts*. https://doi.org/10.17605/OSF.IO/UPJC4

Yuan, S., Fisher, C., Kandhadai, P., & Fernald, A. (2011). You can stipe the pig and nerk the fork: Learning to use verbs to predict nouns. *Proceedings of the 35th annual Boston university conference on language development* (pp. 665–677).

Received September 3, 2023
Revision received July 16, 2024
Accepted August 23, 2024 ■