Received: 23 September 2022 | Revised: 1 March 2023 | Accepted: 4 March 2023

DOI: 10.1111/psyp.14302

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

The P600 as a continuous index of integration effort

Christoph Aurnhammer[1] | **Francesca Delogu**[1] | **Harm Brouwer**[1,2] | **Matthew W. Crocker**[1]

[1]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

[2]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, the Netherlands

Correspondence
Christoph Aurnhammer, Department of Language Science and Technology, Saarland University, Saarbrücken, Germany.
Email: aurnhammer@coli.uni-saarland.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 232722074-SFB 1102

Abstract

The integration of word meaning into an unfolding utterance representation is a core operation of incremental language comprehension. There is considerable debate, however, as to which component of the ERP signal—the N400 or the P600—directly reflects integrative processes, with far reaching consequences for the temporal organization and architecture of the comprehension system. Multi-stream models maintaining the N400 as integration crucially rely on the presence of a semantically attractive plausible alternative interpretation to account for the absence of an N400 effect in response to certain semantic anomalies, as reported in previous studies. The single-stream Retrieval–Integration account posits the P600 as an index of integration, further predicting that its amplitude varies continuously with integrative effort. Here, we directly test these competing hypotheses using a context manipulation design in which a semantically attractive alternative is either available or not, and target word plausibility is varied across three levels. An initial self-paced reading study revealed graded reading times for plausibility, suggesting differential integration effort. A subsequent ERP study showed no N400 differences across conditions, and that P600 amplitude is graded for plausibility. These findings are inconsistent with the interpretation of the N400 as an index of integration, as no N400 effect emerged even in the absence of a semantically attractive alternative. By contrast, the link between plausibility, reading times, and P600 amplitude supports the view that the P600 is a continuous index of integration effort. More generally, our results support a single-stream architecture and eschew the need for multi-stream accounts.

KEYWORDS

EEG, ERPs, language comprehension, N400, P600, psycholinguistics

1 | INTRODUCTION

In electrophysiological studies of language comprehension, the two most salient components of the event-related brain potential (ERP) signal are the N400 and the P600. It is still under debate, however, which of these two components indexes semantic integration—the core operation of compositionally updating an unfolding utterance meaning

representation with incoming information—during online language comprehension. Traditionally, semantic integration has been attributed to the N400 component (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004), such that its amplitude is continuously related to integration effort, a mapping that underpins several contemporary neurocomputational models of comprehension (for a review see Eddine et al., 2022). The P600 has traditionally been discussed in relation to syntactic and structural processing (Hagoort et al., 1993; Osterhout & Holcomb, 1992). This linkage of the N400 to semantic integration and the P600 to purely structural processing is challenged, however, by studies employing semantic role violations, such as "the hearty meal was devouring/devoured" (Kim & Osterhout, 2005, see also Hoeks et al., 2004; Kolk et al., 2003; Kuperberg, 2007; Kuperberg et al., 2003; van Herten et al., 2005, 2006), which lead to P600 rather than N400 effects relative to baseline. To reconcile these "semantic P600" findings with the traditional functional roles of the N400 and the P600, multi-stream models have been proposed which postulate distinct cognitive mechanisms that trigger either an N400 increase or a P600 increase, but typically not both (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for reviews). Motivated by several limitations of these multi-stream models, Retrieval–Integration (RI) theory (Brouwer et al., 2012, 2017) offers an alternative, single-stream account which explains semantic P600 findings by interpreting the N400 as reflecting lexical retrieval (Kutas & Federmeier, 2000, 2011; Lau et al., 2008, 2009; van Berkum, 2009, 2010) and reinterpreting the P600 as a *continuous* index of integration effort. We here employ an experimental design that tests the graded nature of the P600 as an index of integration effort, while also teasing apart the different predictions made by RI theory and multi-stream models about which ERP component should be modulated.

1.1 | Multi-stream models

Multi-stream models typically consist of two processing streams (but see Kuperberg, 2007): a *semantic* stream, linked to the N400, and an *algorithmic* stream linked (indirectly) to the P600. The precise mechanisms thought to underlie these streams vary. The Semantic Attraction account (SA, Kim & Osterhout, 2005), Monitoring Theory (MT, van Herten et al., 2005, 2006), and the extended Argument Dependency Model (eADM, Bornkessel-Schlesewsky & Schlesewsky, 2008), for instance, characterize the semantic stream as assigning thematic roles based on plausibility heuristics and world knowledge, independent of morpho-syntactic cues (see also the Processing Competition account, Kos et al., 2010). In the Continued Combinatory Analysis model (CCA, Kuperberg, 2007), the *semantic memory-based stream* computes semantic features and categorical relationships between words and compares them with pre-existing relations stored in semantic memory. Finally, in a more recent model proposed by Michalon and Baggio (2019), the semantic stream constructs an interpretation of the input by assigning grammatical roles based on lexical–semantic information. While the precise conceptualization of this stream varies across multi-stream models, the absence of an N400 effect in "semantic P600" studies is explained by these accounts in a similar manner: The semantic processing stream is agnostic to the syntactic constraints of the input and thus fails to detect a semantic anomaly whenever a semantically plausible (but syntactically unlicensed) alternative interpretation can be constructed from the content words encountered thus far. In sum, multi-stream accounts typically explain the absence of an N400 effect in semantic P600 findings by positing the presence of a form of semantic attraction (e.g., for the more plausible "the hearty meal was devoured" upon encountering "devouring"; see Li & Ettinger, 2023; Rabovsky et al., 2018; Ryskin et al., 2021 for more recent instantiations of a similar line of reasoning).

The other stream, called *algorithmic stream* (van Herten et al., 2006), *syntactic stream* (Kim & Osterhout, 2005; Kos et al., 2010), or *combinatorial stream* (Kuperberg, 2007), has been described as constructing an interpretation of the input by taking into account morpho-syntactic cues. Again, the conceptualization of this stream changes depending on the specific model. For example, in the eADM model, this stream assigns *thematic* roles based on syntactic "prominence" information (Bornkessel-Schlesewsky & Schlesewsky, 2008). In the CCA, the combinatorial stream combines words based on morpho-syntactic constraints and is complemented with a stream sensitive to semantic–thematic constraints such as animacy (Kuperberg, 2007). In the model proposed by Michalon and Baggio (2019), the syntactic stream assigns grammatical roles based on word position and parts of speech.

Crucially, on these multi-stream models, semantic P600 effects do not directly result from variations in processing cost within the algorithmic stream but rather from situations in which the interpretations generated by the semantic and the algorithmic streams disagree. For example, at the word "devouring," the algorithmic stream assigns the syntactically cued role of *agent* to "meal," which conflicts with the interpretation generated by the semantic stream in which "meal" is the *theme* for "devour." It is this conflict that is posited to result in a P600 effect relative to baseline. Crucially, the absence of an N400 effect together with the presence of a P600 effect for semantic anomalies such as those

induced by implausible thematic roles depends on the availability of a semantically attractive alternative interpretation, for instance one in which the thematic roles are reversed. If such an alternative is not present, multi-stream models predict an N400 increase indexing integration difficulty for the anomalous word in the semantic stream, but no P600 increase, as the outputs of the streams should not be in conflict.

1.2 | Retrieval–Integration theory

Retrieval–Integration theory proposes an alternative, single-stream account in which the N400 is taken to reflect retrieval of word meaning and the P600 is taken to index semantic integration effort (Brouwer et al., 2012, 2017).

Conceptually, RI theory relies on a notion of retrieval that is grounded in the semantic access/retrieval view of the N400 (Kutas & Federmeier, 2000, 2011; Lau et al., 2008, 2009; van Berkum, 2009, 2010), on which semantic/conceptual knowledge associated with a word form—that is, its meaning—is accessed in long-term memory. This retrieval process is cued both by association and by expectation and, indeed, associative and expectation-based influences on retrieval facilitation have been shown to manifest in additive N400 modulations (Aurnhammer et al., 2021). Critically, while associative and expectation-based influences join in facilitating retrieval of word meaning for the current word form, RI theory assumes this process to be non-combinatorial and non-compositional in nature. That is, while the utterance meaning representation influences retrieval of word meaning, the retrieval process itself, as reflected in the N400, does not entail any form of compositional update of the utterance meaning representation.[1] Integrative processes are instead manifest in the P600 component. Conceptually, integration is the updating in working memory of the incrementally constructed utterance meaning representation with the retrieved word meaning. On the RI account, this notion of integration implies a combinatorial process that relies not only on semantic but, critically, also on pragmatic and morpho-syntactic information.

More explicitly, RI theory posits that the word-by-word processing of a sentence is defined by the *process* function (Brouwer et al., 2021):

process (*word form, utterance context*) → *utterance representation*

retrieve (*word form, utterance context*) → *word meaning* [~ N400]

integrate (*word meaning, utterance context*) → *utterance meaning* [~ P600]

Incoming word forms are mapped onto an utterance representation, while taking utterance context, that is, the utterance representation constructed so far, into account. The process function is, however, divided into two subprocesses—*retrieve* and *integrate*—which are linked to the N400 and the P600 component, respectively. The *retrieve* function maps incoming word forms onto a representation of word meaning, while taking utterance context into account. In the neurocomputational model instantiation of the theory (Figure 1), the N400 is taken to be proportional to the distance of the **retrieval** layer at the previous processing step to that at the current processing step. The retrieval process is facilitated—and N400 amplitude attenuated—when the meaning of an incoming word is primed associatively or contextually. The absence of an N400 effect for "the hearty meal was devouring/devoured" is explained by the similar associative priming that both target words receive from the context. Thus, the process underlying the N400 is restricted to accessing word meaning in long-term memory and mapping it into working memory, and extends neither to "quasi-compositional" integration—as proposed by several multi-stream models—nor to compositional integration of word meaning with the utterance meaning representation constructed up to that point, as proposed by the integration view of the N400. The output of the *retrieve* function serves as an input to the *integrate* function, which maps the retrieved word meaning onto an updated utterance meaning representation while taking previous utterance context into account. The P600 is taken to proportionally reflect the distance in activation between the **integration** layer at the previous processing step and that at the current processing step. The P600 increase for "devouring" compared to "devoured" thus results from a more difficult integration process due to the implausibility of *meal* fulfilling the agent role.

The interpretation of the P600 as an index of integration effort is, however, not limited to role-reversal

[1]This perspective on retrieval separates RI theory from the hybrid view of the N400. On RI theory, retrieval is taken to include both what has, on the hybrid view, been called preactivation—the process by which "the semantics of the context activates lexical features of an incoming word" (Baggio & Hagoort, 2011, p. 1348) and the process by which "different sources of information converge on a common memory representation" (Baggio & Hagoort, 2011, p. 1347, the hybrid view calls the latter notion "integration" and does not posit this process to be reflected in the N400). RI theory diverges from the hybrid view, in that the latter additionally posits unification—the "integration of word meaning into an unfolding representation of the preceding context" (Hagoort et al., 2009, p. 1)—to be indexed by the N400. This update is what RI theory calls integration and attributes to the P600.

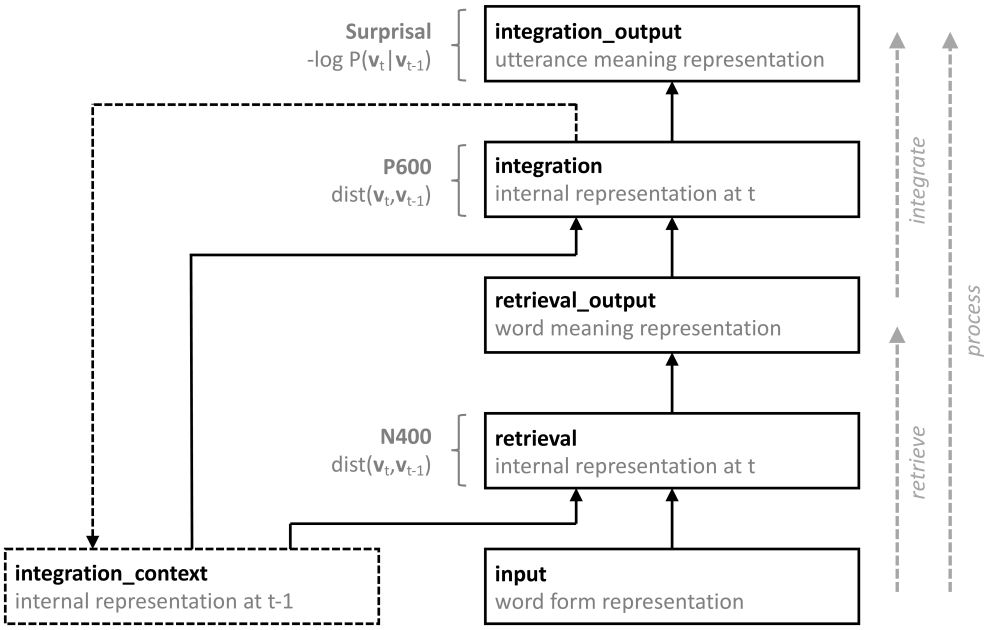

FIGURE 1 Schematic architecture of the neurocomputational instantiation of Retrieval–Integration theory, implementing word-by-word language processing through the *retrieve* and *integrate* functions. For full detail on the model implementation, see Brouwer et al. (2021).

manipulations but naturally extends to those semantic P600 findings induced not only by semantic and pragmatic factors (Burkhardt, 2006, 2007; Cohn & Kutas, 2015; Delogu et al., 2019; Dimitrova et al., 2012; Hoeks et al., 2013; Regel et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Xu & Zhou, 2016) but also those induced by manipulations of syntax (Gouvea et al., 2010; see Brouwer et al., 2012; Delogu et al., 2019 for discussion) and syntax-driven semantic composition (Fritz & Baggio, 2020, 2022). Importantly, on the RI account, the amplitude of the P600 should not be a binary response to violating stimuli but should rather be sensitive to integration effort on a continuous scale (Brouwer et al., 2012), reflecting comprehension-centric surprisal (Brouwer et al., 2021). Preliminary evidence for this prediction has been presented in a post hoc analysis by Aurnhammer et al. (2021), who demonstrated a graded response of both the N400 and the P600 to congruous sentences that varied in target word expectancy.

Crucially, the notion of integration assumed by RI theory is not coextensive with the aspects of integration proposed for the semantic stream by multi-stream models. Rather, integration in the RI model is closer to the algorithmic stream, in that integration is posited as morpho-syntactically constrained utterance meaning composition. Importantly, however, while most multi-stream models do not directly attribute any electrophysiological processing correlate to the algorithmic stream, RI theory takes the P600 to be directly proportional to the change in utterance meaning representation induced by the current word meaning.

1.3 | Disentangling multi-stream models and RI theory

While both multi-stream models and RI theory are able to account for semantic P600 effects elicited in the presence of semantic attraction (e.g., caused by role reversals), they differ in predicting which component should reveal integrative effort *in the absence* of a semantically attractive alternative interpretation. As previously discussed, multi-stream models predict an N400 effect reflecting an unrepairable semantic anomaly and no P600 effect, as no conflict should arise between the semantic and the algorithmic stream, relative to a plausible baseline. By contrast, the RI account predicts the N400 to be modulated by the degree to which the meaning of the implausible word is associatively primed and contextually expected, and a P600 effect reflecting continuous semantic integration effort, relative to a plausible baseline.

1.3.1 | Semantic P600 effects in a wider discourse

Here, we present an experimental design that directly tests the predictions of multi-stream models against those of RI theory. To this end, we build on the design employed by Nieuwland and van Berkum (2005) in which a context paragraph is followed by a critical region including either a plausible (coherent: "the woman told the tourist") or an implausible (incoherent: "the woman told the suitcase")

TABLE 1 Experimental stimulus from the design of Nieuwland and van Berkum (2005), translated from Dutch.

Introduction	
A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos.	
Coherent continuation	*Incoherent continuation*
Next, the woman told the *tourist* that she thought he looked really trendy. The tourist grabbed the woman's hand and eagerly asked her for a date. But the woman reprimanded the tourist for being pushy and told him to just get on the plane right away.	Next, the woman told the *suitcase* that she thought he looked really trendy. The suitcase grabbed the woman's hand and eagerly asked her for a date. But the woman reprimanded the suitcase for being pushy and told him to just get on the plane right away.

Note: Underlines added by the authors of this article.

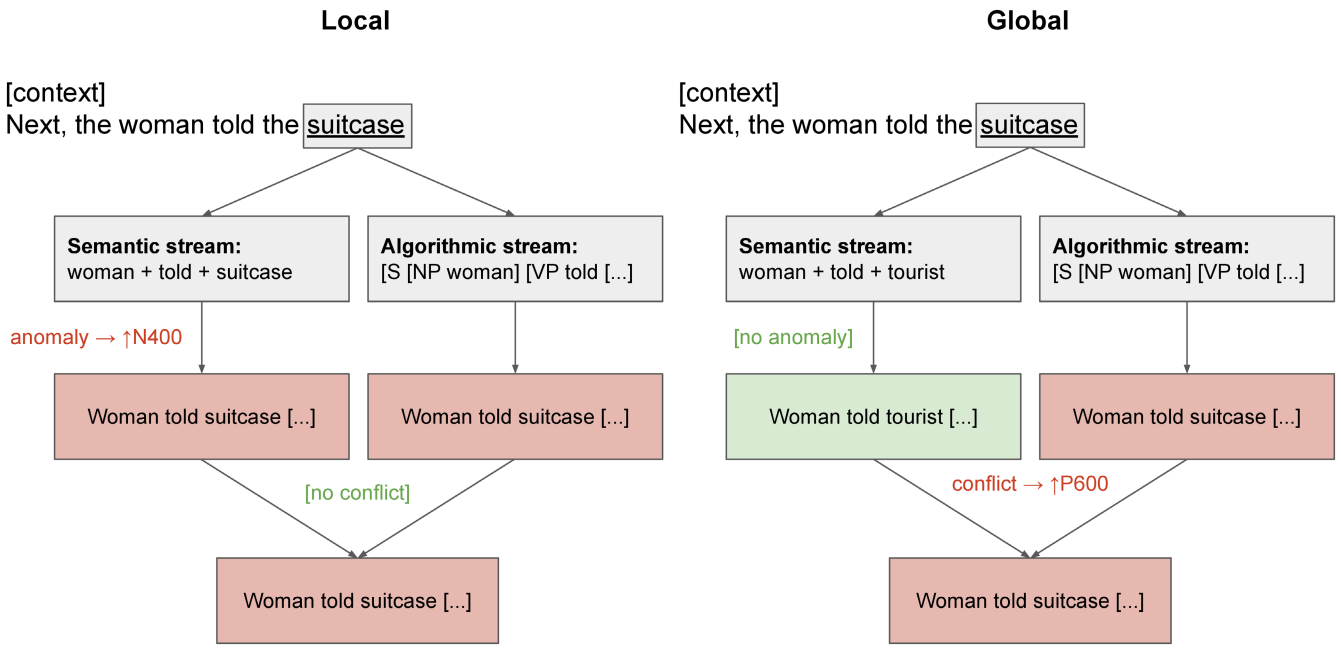

FIGURE 2 Schematic overview of multi-stream explanations assuming either a *local* or a *global* revision mechanism.

target word (Table 1). Crucially, both target words, "tourist" and "suitcase," are mentioned several times in the preceding context paragraph. Stimuli were presented in spoken form and without a task. The contrast of the implausible (incoherent) "suitcase" to the plausible (incoherent) "tourist" elicited a broadly distributed P600 effect, but no N400 effect.

This result seems inconsistent with multi-stream accounts: When encountering the implausible target word "suitcase," there is no *locally* available semantically attractive alternative—for example, through sentence-internal permutation of thematic roles and/or morphological inflection—that would yield a plausible interpretation of the sentence. As a result, multi-stream models predict an N400 effect, reflecting the difficulty in arriving at a semantically plausible analysis when compared to a plausible sentence, but no P600 effect, as there is no disagreement between the independent semantic stream and the algorithmic stream (see Brouwer et al., 2012 for discussion; a schematic multi-stream analysis is given in Figure 2, left). It has been argued, however, that a semantically attractive alternative may be *globally* available in the larger discourse (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007, for discussion). That is, as both "tourist" and "suitcase" are salient entities in the discourse, which have been mentioned numerous times, the interpretation of the coherent condition ("the woman told the tourist") may actually be a strong attractor in the incongruent condition. In other words, the salience of the plausible noun phrase "the tourist" may *distract* the system away from the actual noun phrase "the suitcase." If this is the case, a multi-stream account of this result would entail the independent semantic stream encountering no difficulty

in producing a plausible analysis, which should lead to no N400 modulation, thereby yielding a conflict with the algorithmic processing stream (which arrives at the analysis "the woman told the suitcase"), thereby triggering a P600 effect relative to baseline (see Figure 2, right).

Retrieval–Integration theory attributes the absence of an N400 effect to facilitated retrieval. That is, the lexical repetition of both the congruent and incongruent target words leads to maximal priming of their meaning. Indeed, in line with this interpretation, the N400 effect resurfaced, for similar stimuli presented in story-initial position, that is, without any preceding context mentioning the target words (see figure 4 in Nieuwland & van Berkum, 2005), due to the absence of equal priming for "suitcase" and "tourist."[2] The presence of a P600 effect, in turn, reflects the difficulty in integrating "suitcase" versus "tourist" in "the woman told [...]," as the former yields an interpretation that goes against world knowledge. If we accept the independent semantic processing stream of multi-stream models to be able to compute a *globally* available semantically attractive alternative interpretation, then multi-stream models and RI theory make the same N400 and P600 predictions, and both account for the Nieuwland and van Berkum (2005) data. Crucially, however, if no such alternative interpretation is available, the accounts make diverging predictions: Multi-stream models predict an N400 effect and no P600 effect, while RI theory predicts no N400 effect and a P600 effect relative to baseline. Furthermore, while previous studies observing semantic P600 effects typically employed binary designs, RI theory makes the specific prediction that P600 amplitude should be a function of graded integration effort (see Aurnhammer et al., 2021, for preliminary support). To test these diverging predictions, we here present an adapted version of the Nieuwland and van Berkum (2005) design.

1.3.2 | Global attraction versus continuous integration

The adapted design implements several manipulations (see Table 2). First, we created a baseline condition, in which the target word is expected and plausible and no processing difficulties should ensue (Condition A). In order to test the prediction of multi-stream models that it is the availability of a semantically attractive alternative that explains the absence of an N400 effect and the presence of a P600 effect, we constructed one condition such that an alternative is made *globally* available by a distractor word in the context (Condition B). In another condition, no such alternative is available (Condition C) and we compare both conditions to the unmanipulated baseline (Condition A). Furthermore, to test for the gradedness of integration effort, the target word in Condition B has intermediate plausibility, in that it renders the interpretation semantically unlikely yet possible, while Condition C is implausible, yielding a semantic anomaly (see Table 4 for more examples). Finally, to maximize comparability of target word processing across conditions, our design employs a context rather than a target manipulation design and we harness lexical repetition to maximally and equally prime the target words in the three conditions.

In the adapted design, multi-stream models predict a P600 and no N400 effect for Condition B relative to Condition A (see Table 3). This is because the anomaly is repairable by replacing the anomalous interpretation resulting from the observed word with the *globally* available alternative interpretation that derives from the distractor

TABLE 2 Experimental design of the present study.

Context
Ein Tourist wollte seinen riesigen **Koffer** mit in das Flugzeug nehmen. Der **Koffer** war allerdings so schwer, dass die Dame am Check-in entschied, dem Touristen eine extra Gebühr zu berechnen. Daraufhin öffnete der Tourist seinen **Koffer** und warf einige Sachen hinaus. Somit wog der **Koffer** des einfallsreichen Touristen weniger als das Maximum von 30 Kilogramm.
*A tourist wanted to take his huge **suitcase** onto the airplane. The **suitcase** was however so heavy that the woman at the check-in decided to charge the tourist an extra fee. After that, the tourist opened his **suitcase** and threw several things out. Now, the **suitcase** of the ingenious tourist weighed less than the maximum of 30 kilograms.*
Condition A: Plausible, baseline
Dann verabschiedete die Dame den Touristen und danach ging er zum Gate.
Then dismissed the lady the tourist and afterwards he went to the gate.
Condition B: Less plausible, attraction
Dann wog die Dame den Touristen und danach ging er zum Gate.
Then weighed the lady the tourist and afterwards he went to the gate.
Condition C: Implausible, no attraction
Dann unterschrieb die Dame den Touristen und danach ging er zum Gate.
Then signed the lady the tourist and afterwards he went to the gate.

Note: German word order is preserved for the English transliterations of the final sentences. Target words are underlined and distractor words are highlighted in boldface.

[2]Visual inspection suggests that this N400 effect co-occurs with an increase in P600 amplitude.

TABLE 3 N400 and P600 predictions of multi-stream models and Retrieval–Integration theory for the current design.

	Multi-stream		Retrieval–Integration	
	N400	**P600**	**N400**	**P600**
A: Plausible, no attraction	–	–	–	–
B: Less plausible, attraction	–	+	–	+
C: Implausible, no attraction	+	–	–	++

TABLE 4 Four example items, transliterated from German.

Item 2
A teacher saw an old world map in the showcase of an antique shop. Such an authentic artifact appeared suitable for his classroom and he approached the **saleswoman** ...
A: Then bought the teacher the map ...
B: Then kissed the teacher the map ...
C: Then filled the teacher the map ...
Item 4
While building a table, a **carpenter** broke his nice hammer into pieces ...
A: Then took the apprentice the hammer ...
B: Then sneered-at the apprentice the hammer ...
C: Then ate the apprentice the hammer ...
Item 11
In a foreign city, a vacationer booked a guided tour. The **guide** was happy that the vacationer was interested and gifted him a flyer ...
A: After the tour folded the vacationer the flyer ...
B: After the tour commended the vacationer the flyer ...
C: After the tour cooked the vacationer the flyer ...
Item 18
A young lady wanted to have a **jewel** evaluated by a jeweler ...
A: Delighted remunerated the lady the jeweler ...
B: Delighted marveled-at the lady the jeweler ...
C: Delighted seasoned the lady the jeweler ...

Note: Target words are underlined, distractor words are highlighted in boldface.

word, similar to the original study. In Condition C, however, no such alternative interpretation is licensed by the context and hence multi-stream models predict an N400 effect and, critically, no P600 effect relative to the baseline condition. RI theory predicts that no N400 differences should be produced across conditions due to the lexical repetition of the target word in the context paragraph, maximally facilitating lexical retrieval of its meaning. Under the hypothesis that P600 amplitude continuously indexes the effort of integrating word meaning with the utterance meaning representation constructed so far, the P600 is predicted to be graded for plausibility with increasing amplitude for conditions $A < B < C$. In sum, while multi-stream models predict a P600 effect for Condition B and an N400 effect for Condition C relative to the baseline Condition A, RI theory predicts the absence of N400 effects, and graded P600 amplitude differences across conditions.

On the assumption that reading times provide an index of overall word-by-word processing effort, we first collected self-paced reading time data for our novel design. We expect that reading times should be graded for target word plausibility, reflecting graded integration effort. Subsequently, we recorded event-related potentials for the same stimuli, allowing for a direct comparison between behavioral and neurophysiological indices of integrative processing effort (see Brouwer et al., 2021, for discussion).

2 | EXPERIMENT 1: SELF-PACED READING

2.1 | Method

Code and data required to reproduce the analyses are made publicly available.[3] All studies were conducted with ethics approval of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS).

2.1.1 | Materials

The materials were optimized to be used in the same form in the self-paced reading study and the electroencephalography (EEG) study (see Appendix S1 for the full list of German stimuli). In the creation of the stimuli, we translated and adapted items from Nieuwland and van Berkum (2005) where possible, and otherwise developed new items. In total, we developed 96 items for which we changed the original target manipulation to a context manipulation design. Employing a context manipulation design in which the target word is the same across conditions is intended to reduce effects due to differences in word length, frequency, etc. Every item had the same context paragraph in each condition.

The context paragraph repeatedly mentioned both the target word as well as a distractor word. The target word and the distractor word were mentioned the same amount of times within item (three or four times). Presenting the target word several times in the

[3] https://github.com/caurnhammer/psyp23rerps.

	Cloze			Plausibility		
Condition	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Target						
A	0.80	0.20	0.33–1.00	5.84	0.93	3.60–7.00
B	0.09	0.11	0.00–0.40	3.69	1.33	1.50–6.30
C	0.02	0.04	0.00–0.20	1.42	0.33	1.00–2.40
Distractor						
A	0.05	0.90	0.00–0.33	2.53	1.34	1.10–6.30
B	0.78	0.17	0.33–1.00	5.94	1.05	2.40–7.00
C	0.03	0.06	0.00–0.20	1.66	0.69	1.00–4.80

TABLE 5 Averages, standard deviations, and ranges for the results of two norming studies that collected cloze probabilities and seven-point scale plausibility ratings for the target and the distractor word.

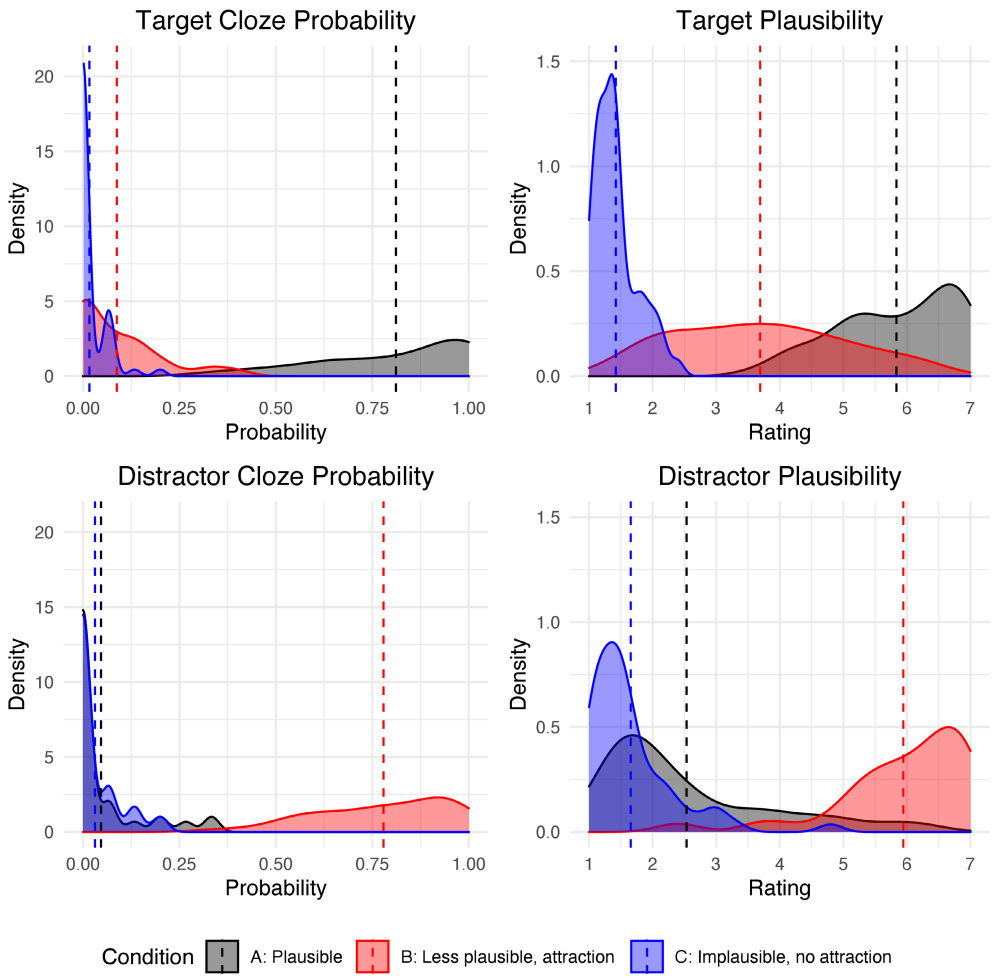

FIGURE 3 Densities for the results of two norming studies that collected cloze probabilities and seven-point scale plausibility ratings for the target and the distractor words. Vertical lines indicate per-condition averages.

context paragraph should maximally prime the target word's meaning, when presented in target position. Under RI theory, we thus expect no N400 (retrieval) effect across conditions (see Brouwer & Crocker, 2017; Brouwer et al., 2012). Which of the two words—target or distractor—was last mentioned in the context was approximately balanced across items.

The context paragraph was followed by a manipulated final sentence. Conditions differed only in the main verb of the final sentence, rendering the target word of the sentence—that is, the direct object—plausible (Condition A, "the lady *dismissed* the tourist"), less plausible (Condition B, "the lady *weighed* the tourist"), or implausible (Condition C, "the lady *signed* the

tourist"). Indeed, Condition C creates a standard semantic anomaly by violating the selectional restrictions of the main verb. The only important difference to a standard semantic anomaly is that the target word has been presented several times before appearing in target position.[4] Taken together, this allows us to assess whether plausibility results in graded effects on both RTs and the P600. Additionally, the distractor word, which is never presented in target position, was either expected (Condition B, "the lady *weighed*" attracting "suitcase"), or not expected (Condition A and C), allowing us to investigate whether the presence of a semantically attractive alternative interpretation modulates the presence of P600 (Condition B; semantic attraction) or N400 effects (Condition C; no semantic attraction) in the ERP experiment. The final sentence of each item ended with an additional clause following the target word ("[...] and afterwards he went to the gate"), which avoids placement of the target in sentence-final position and allows us to capture spillover effects in reading times. Table 4 shows four more transliterated items.

Cloze

We collected cloze probabilities to validate the differential expectancy of both the target and distractor word across conditions. Sentence completions were collected in a web-based experiment using the software PCIbex (Zehr & Schwarz, 2018), which we also used for all other web-based norming studies and experiments reported here. We did not use filler items, since the materials up to the target word do not contain any anomalies. Participants were presented with the entire context paragraph and the final sentence up to—but not including—the determiner of the target word. That is, we did not provide a determiner as the grammatical gender of German would constrain the set of possible completions.

In order to maximize the contrast of high expectation for the target (Condition A) or the distractor word (Condition B), we obtained Cloze probabilities in two rounds. Sentence contexts for implausible words were created such that they do not raise strong expectations for any specific word (Condition C). In total, we collected responses from 90 participants, who were recruited through Prolific Academic Ltd. and each was paid £7.50. We selected the 60 best items based on the results of the cloze task. Alternative cloze probabilities for any other word in Condition C were kept below 0.27 (mean = 0.20; $SD = 0.07$). The resulting cloze probabilities for the target and distractor word across the three conditions are presented in Table 5 and Figure 3 (left). Target word cloze probability is high in Condition A, indicating high expectancy of the target word in the baseline condition, which should therefore induce only low integrative effort. In Condition B, participants actively produced the distractor word rather than the target word, indicating that the distractor word indeed makes a semantically attractive alternative interpretation (globally) available in this condition. In Condition C, expectancy of both the target word and the distractor word was low. The latter suggests that the alternative interpretation available for Condition B is removed in Condition C. In sum, the cloze probabilities suggest that the availability of the semantically attractive alternative interpretation has been manipulated successfully (Condition A: baseline; Condition B: semantic attraction; Condition C: no semantic attraction). We turn to a second norming study in which we collect plausibility ratings to discern whether the target words of conditions B and C—which were similarly unexpected—indeed differ in their plausibility.

Plausibility

In a second norming study, we collected plausibility ratings for the target and distractor words on a seven-point Likert scale, 7 indicating "very plausible" and 1 indicating "not plausible." In total, 60 participants were recruited through Prolific Academic Ltd., and each was paid £7.50. For the rating task, the final sentence was presented in one paragraph together with the context material, with the aim to ensure reading of the entire paragraph and not only the final sentence. Participants were instructed to rate the plausibility of the final sentence in light of the context. We excluded the final sentence continuation ("and afterwards he went to the gate") to maximize rating the target word rather than another part of the final sentence. During the rating task, there were 10 items with attention checks which presented mid-paragraph instructions to rate this trial with a given number (either 1 or 7). On average, participants completed 98 % of attention checks successfully (mean = 98.19 %; $SD = 4.09$; range 83.33–100.00 %). The resulting plausibility ratings are reported in Table 5 and Figure 3 (right). Target word plausibility is stepped across conditions (A > B > C), which should result in a similarly graded effect of integration effort on the target in the three contexts. Distractor word plausibility is high in Condition B while in Condition A and C, distractor word plausibility is low, again supporting the availability of a semantically attractive alternative interpretation in Condition B.

[4]Furthermore, most of these semantic anomalies render reference transfer to a related entity unlikely. For instance, while it is conceivable that reference may be transferred from "tourist" to the "tourist's ticket" in the example stimulus, for most of our stimuli, no such reference transfer is licensed (e.g., "the apprentice ate the hammer").

TABLE 6 Correlations between cloze probabilities and plausibility ratings of the target and distractor words.

	Cloze		Plausibility	
	Target	**Distractor**	**Target**	**Distractor**
Cloze				
Target	1.00	−0.40	0.79	−0.24
Distractor	−0.40	1.00	0.01	0.88
Plausibility				
Target	0.79	0.01	1.00	0.22
Distractor	−0.24	0.88	0.22	1.00

Correlations between target and distractor word cloze probability and plausibility are reported in Table 6. Our analyses will focus on target word plausibility to investigate graded effects of plausibility and on distractor cloze to investigate additional effects of semantic attraction. As the correlations show, these predictors are effectively independent ($r = 0.01$).

2.1.2 | Participants

Forty-three participants were recruited through Prolific Academic Ltd., to take part in the web-based self-paced reading experiment. One participant was excluded due to inattentive reading, as shown by low accuracy on the task (60% correct; see below for specifics of the task). The remaining 42 participants (mean age 24.43; SD 3.7; age range 18–32; 15 male, 27 female) were all native speakers of German (two early bilinguals) and had not indicated any language-related disorders or literacy difficulties. They did not participate in any other studies reported in this article. All participants gave their consent by agreeing to a consent form and were paid £7.50 for their participation.

2.1.3 | Procedure

We conducted the self-paced reading experiment as a web-based study. On each trial, participants were prompted to press the Enter key to start, after which they were presented with a context paragraph. Upon pressing the Enter key again, a hash sign was presented centrally, indicating the position of the words of the final sentence. From here on, participants pressed the Space bar to proceed to the next word, each presented centrally. After three practice items, stimuli were presented in three blocks with 35 items each, summing to a total of 105 items, 45 of which were fillers. For half of the participants, the blocks and the items within them were presented in reverse order. On 46% of trials—half of the experimental trials and on two-fifths of the fillers—participants were presented with a comprehension question to which they had to answer either *Yes* or *No* (mapped to the *D* and *K* keys). Comprehension questions had *Yes* and *No* as correct answer on 50% of the questions and they could concern the context paragraph or the final sentence, within which they could focus on the manipulated region or the final sentence completion. To encourage attentive reading, we provided coarse feedback on participants' response accuracy after the practice session and after each block. Participants were encouraged to take a short break between blocks.

2.1.4 | Analysis

We excluded trials if reading time on any critical region was lower than 50 ms or higher than 2500 ms and if reaction time on the task (if there was one on that trial) was lower than 50 ms or higher than 6,000 ms. Based on these criteria, 47 of 2520 trials were excluded (1.87%). All results and analyses reported below are computed after exclusion.

Log-transformed reading times were analyzed with a linear mixed effects regression re-estimation technique (cf. Aurnhammer et al., 2021), using the MixedModels package for Julia (Bezanson et al., 2017). This technique fits reading time models separately on each region of interest, allowing to trace across regions the relative influence and significance of each predictor in the regression equation as well as the residual error, that is, the difference between the observed data and the forward estimates computed by the models. As predictors of interest, we focus on target word plausibility and distractor cloze probability. Plausibility ratings will serve as a continuous predictor to operationalize integration difficulty of the target word. Distractor cloze probability serves as a predictor that will explain any additional effort incurred by the availability of a semantically attractive alternative interpretation. Random intercepts as well as random slopes for each predictor are estimated for both subjects and items. The full model specification is

$$Y = \beta_0 + S_0 + I_0 + (\beta_1 + S_1 + I_1)\text{Plaus} + (\beta_2 + S_2 + I_2)\text{Clozedist} + \epsilon \quad (1)$$

in which β_0 represents the fixed-effect intercept and β_1 and β_2 refer to the fixed-effect coefficients of plausibility and distractor cloze probability. The S and I terms represent random intercepts and slopes for subjects and items. The unexplained variance in the data is represented by the residual error term ϵ. All predictors were standardized, centering their average value on zero and expressing them on a scale of standard deviations. Standardizing predictors

TABLE 7 Task performance on the comprehension questions in the self-paced reading experiment.

	Accuracy			Reaction time		
Condition	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
A	96.7%	6.1%	80.0–100.0%	2900 ms	560 ms	1566–3820 ms
B	95.3%	8.3%	70.0–100.0%	3032 ms	567 ms	1986–4106 ms
C	96.0%	7.0%	77.8–100.0%	3047 ms	586 ms	2086–4259 ms

Note: Accuracy and reaction times were computed across subjects.

additionally has the effect that the intercept will equal the mean of the data to which the model is fitted. Plausibility was also inverted, as we predict that higher reading times ensue for lower plausibility ratings. We run separate analyses for the different regions of interest, which we treat as separate families of hypotheses. Hence, we do not correct for multiple comparisons.

2.2 | Results

2.2.1 | Comprehension questions

Participants answered comprehension questions on half of the experimental items. Descriptive metrics for accuracy and reaction times were computed across subjects. Average accuracy was 96.8% ($SD = 5.2$, range = 80–100.0%). Mean reaction time was 3098 ms ($SD = 619$, range = 1907–4426 ms). Accuracies and reaction times per condition are given in Table 7.

2.2.2 | Reading times

Figure 4 displays log-transformed reading times, split up per condition, on the Pre-critical region (the ambiguous article "den"/"the" of the target word), the Critical region (the target word "tourist"), the Spillover region ("and"), and the Post-spillover region ("afterwards"). Visual inspection of the data suggests that already on the Pre-critical and Critical regions, Condition C is read slower than Conditions A and B. On the Spillover region, Condition B and C are slowed down. Lastly, on the post-spillover region, reading times appear to pattern with the three levels of Conditions A, B, and C.

We modeled the reading times as a function of both target word plausibility and distractor cloze probability on each region separately. Figure 5 displays the estimated reading times from these models as well as the residual error, that is, the difference between the observed and the estimated reading times. Visual inspection suggests that the models capture the effect structure in the observed data as evidenced by small residual error across regions and conditions.

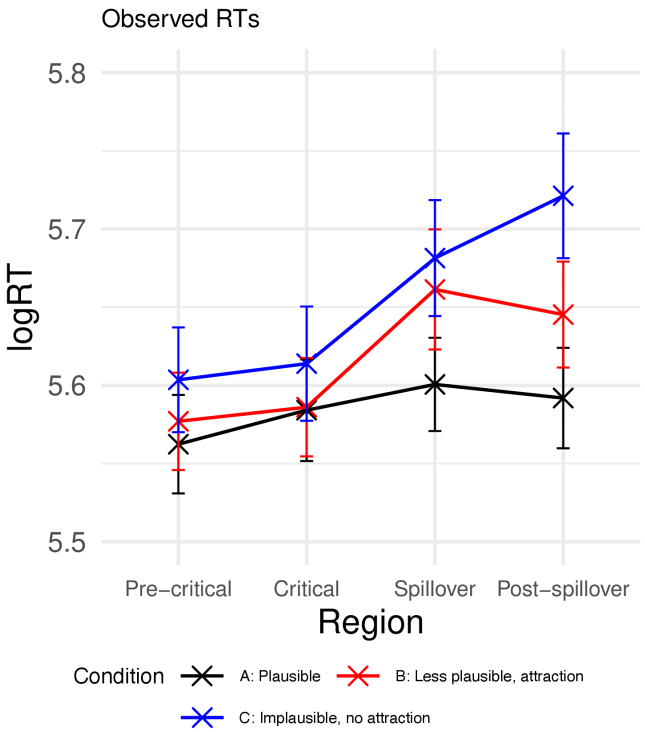

FIGURE 4 Log reading times, averaged per condition from the per-subject averages, on the Pre-critical, Critical, Spillover, and Post-spillover region. Error bars indicate the standard error computed from the per-subject per-condition averages.

Figure 6 (left) displays model coefficients, added to their intercept, for plausibility and distractor cloze probability together with their respective z and p values (right). The positive coefficients for plausibility indicate that lower plausibility predicts slower reading. The coefficient for distractor cloze probability is smaller and changes sign moving from the Critical to the Spillover and to the Post-spillover region, indicating that this predictor estimates slower or faster reading time depending on the region of interest. The z and p values demonstrate that target word plausibility significantly predicts reading times across all regions, interestingly also including the Pre-critical region, while no significant contribution of distractor cloze probability was found.

Reading in the implausible Condition C is slowed already prior to the target word, presumably due to differences in processing of the main verbs preceding the targets. This raises the question to what extent reading time

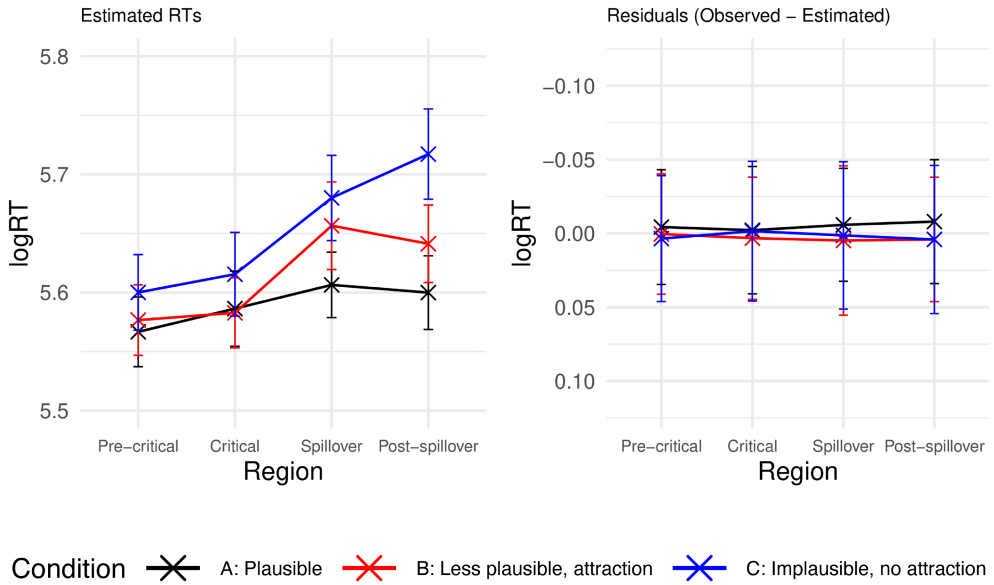

FIGURE 5 Estimated log-reading times (left) and residual error (right), averaged per condition, on the Pre-critical, Critical, Spillover, and Post-spillover regions.

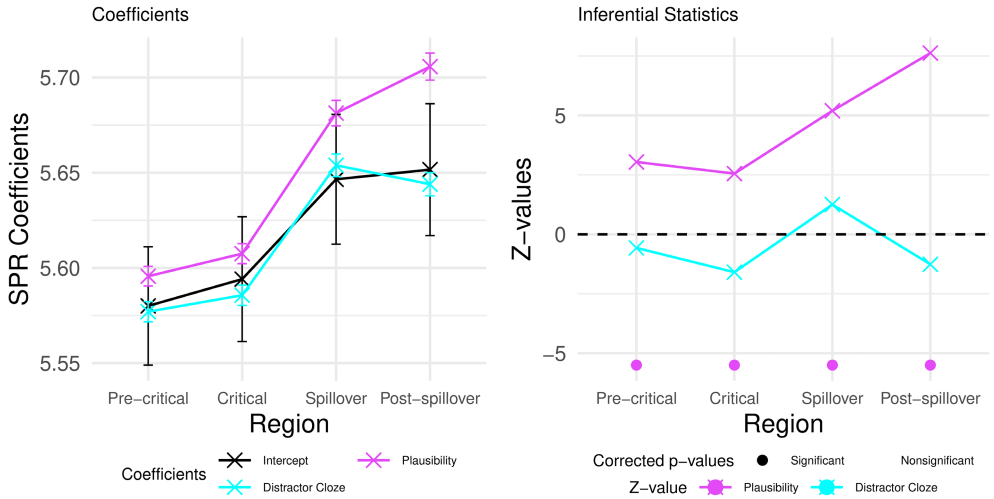

FIGURE 6 Coefficients (left; added to their intercept), effect sizes (z values), and p values (right). Error bars indicate the standard error of the coefficients in the fitted statistical models.

differences observed on and after the critical word are due to the plausibility of the target word itself, rather than due to the different contexts. To answer this question, we included the reading time on the Pre-critical region as a predictor into our analyses, allowing the models to capture any pre-critical reading time offsets. We only z-scored but did not log-transform the Pre-critical RT predictor, in order to avoid identity of the dependent (logRT) and one of the independent variables (Pre-critical RT) on the Pre-critical region. The remaining predictors now explain any systematic variability in reading time over and above reading time offsets present at the Pre-critical region. The resulting coefficients and z values indicate that target word plausibility significantly predicts slowed reading time at the Spillover and Post-spillover regions, over and above what is explained by Pre-critical reading time, whereas the plausibility predictor is no longer significant on the Pre-critical and Critical regions. Distractor cloze probability still does not significantly predict reading times on any region (Figure 7).

2.3 | Discussion

The results of the self-paced reading experiment show that reading times scale gradually with plausibility,

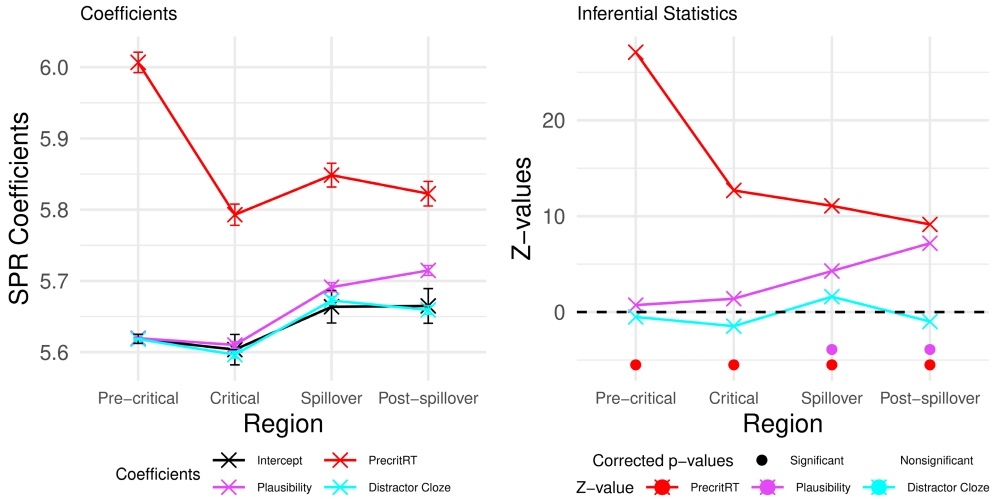

FIGURE 7 Coefficients (left; added to their intercept), effect sizes (z values), and p values (right) from models including Pre-critical reading time as a predictor. Error bars indicate the standard error of the coefficients in the fitted statistical models.

indicating that our manipulation of target plausibility indeed resulted in a graded modulation of integration effort. Furthermore, the regression-based analysis revealed that plausibility is a continuous predictor of reading time.

Based on the traditional surprisal literature (Frank et al., 2015; Levy, 2008; Monsalve et al., 2012), it could be expected that the same items that show modulations in reading times would also elicit a graded N400 response. However, the hypothesis that the P600 reflects integration effort predicts a strong link between this component and late reading time measures (Brouwer & Crocker, 2017; Brouwer et al., 2012). Empirical evidence in support of this is provided by Brouwer et al. (2021) and Aurnhammer et al. (2021), showing that reading time modulations pattern with P600 effects.[5] The obtained reading times thus offer an opportunity to investigate whether the experimental design will result in a graded N400 or P600 pattern.

The current results did not reveal significant reading time modulations due to distractor cloze probability. Hence, our results indicate no significant reading time modulation that can be attributed to the presence of a semantically attractive alternative interpretation in Condition B. However, multi-stream models typically do not make predictions for behavioral measures, and hence, we will not rely on this result to argue against these accounts. Our manipulation does, however, create a prediction disconfirmation, since in the context "Then weighed the lady", the expected word "suitcase" is not presented, while "tourist" is provided instead. Previous research on prediction error cost has not found disconfirmation effects in the behavioral domain using eye-tracking (Frisson et al., 2017; Luke & Christianson, 2016) or self-paced reading (Rich & Harris, 2021). In a self-paced reading experiment by van Berkum et al. (2005), a disconfirmation effect was observed—however, its timing did not coincide with the ERP deflection found for the same stimuli. Similarly, lexical decision times did not exhibit facilitation effects for unrelated, unexpected words in high constraint sentences relative to the same words in low constraint sentences (Schwanenflugel & LaCount, 1988). In line with this previous research, our results suggest that reading times may not be sensitive to unfulfilled expectations. With regard to the comparison of multi-stream models and RI theory, the absence of a significant contribution of semantic attraction (distractor cloze probability) in behavioral measures raises the question whether semantic attraction will modulate the presence of P600 and N400 effects in the ERP signal.

[5]Additionally, effects of association, which were also reflected in N400 amplitude, modulated reading times on the first Spillover region of Aurnhammer et al. (2021). As the current design maximally primes the targets across all conditions, no such association-related effects were expected in the current data.

3 | EXPERIMENT 2: ELECTROENCEPHALOGRAPHY

3.1 | Method

3.1.1 | Materials

The materials were the same as in the self-paced reading experiment (see Section 2.1.1).

3.1.2 | Participants

We recruited 33 participants at Saarland University to take part in the experiment. Three participants were excluded due to excessive eye movement artifacts. The

final 30 participants (mean age 25; $SD = 3.35$; range 18–32; 25 female, 5 male) were right-handed, native speakers of German (six early bilinguals) and had normal or corrected-to-normal vision. None reported any form of color blindness. Participants gave informed, written consent and were paid 25€.

3.1.3 | Procedure

We recorded the EEG while the participants were seated in an electromagnetically shielded, soundproof, and dimly lit chamber. The experiment was presented using E-prime 3 (Schneider et al., 2002). We first presented three practice items, two of which included a comprehension question. Practice items varied in their degree of plausibility. The practice session was followed by three blocks, each containing 35 items, including the same fillers that were used in the self-paced reading experiment. Participants took a break between blocks. Items were presented in pseudorandomized order. For half of the participants, the blocks and the items within them were presented in reverse order. On each trial, participants used a button-box to start the item and were presented with the entire context paragraph which remained on the screen until the button was pressed again. Then, a fixation cross appeared in the center of the screen for 750 ms. After that, the final sentence was presented using rapid serial visual presentation (RSVP). Each word of the final sentence was presented centrally for 350 ms with a 150 ms inter-stimulus interval. If the item contained a comprehension question, the question appeared after the last word of the final sentence. Questions were answered using two buttons that mapped to *Yes*/*No*, highlighted on the screen in green and red color, respectively. The position of the correct and incorrect button varied randomly in order to avoid motor preparation effects.

3.1.4 | Electrophysiological recording and processing

The EEG was recorded using 26 active Ag/AgCl electrodes, positioned on the scalp following the standard 10–20 system. During recording, FCz was used as online reference and AFz as ground. Data were digitized at a sampling rate of 1000 Hz, leading to a temporal resolution at 1 ms increments. Eye movement artifacts were monitored through the electrooculogram of two electrodes placed horizontally at the outer canthi of each eye and two electrodes placed vertically above and below the left eye. We aimed to keep impedances below 5 $k\Omega$ on scalp electrodes and below 10 $k\Omega$ on eye electrodes and did not apply online filtering. We re-referenced the EEG offline to the averages of the left and right mastoid electrodes and band-pass filtered the data between 0.01 Hz and 30 Hz. Epochs ranging from −200 to 1200 ms relative to target word onset were extracted from the EEG signal. Trials with ocular and muscular artifacts were excluded using a semi-automatic procedure. Baseline correction was performed using the 200 ms pre-stimulus interval.

3.1.5 | Analysis

To analyze the data, we apply rERPs (Smith & Kutas, 2015), a regression-based ERP (re-)estimation technique (implemented in Julia; Bezanson et al., 2017), similar to the analysis used for the self-paced reading data. For this analysis, we apply linear regression, as opposed to linear mixed-effects regression, as the analytical solution of solving least-squares regression will provide stable models and faster computation speed. This will allow us to re-estimate the data on all electrodes and to inspect topographic differences in the analyses. In particular, rERPs apply within-subjects regression and the models' parameters and forward solutions are averaged across subjects, analogous to the traditional ERP averaging procedure in which condition averages are computed from the means of individual subjects. The advantage of the rERP technique compared to traditional statistical analyses is that it allows us to gauge the relative explanatory power of target word plausibility and distractor cloze probability across time and electrodes: By computing a separate regression model for each subject on each electrode and time sample, we can trace predictor coefficients, inspect estimated waveforms and residual error, and obtain effect sizes across the temporal and spatial dimensions of the ERP signal. Crucially, this approach goes beyond simple condition contrasts, as we are interested in the continuous relationship between stimulus properties and ERPs. In fact, the rERP analyses themselves are only informed by the continuous by-trial stimulus properties and not by any explicit condition coding. That is, we only average by condition after fitting the models, to assess the extent to which our predictors capture the effect structure across conditions.

We will apply the same predictor combination that we used for the analysis of the reading times and model the ERP signal as a function of target word plausibility and distractor cloze probability. The model specification for the rERP models is

$$Y = \beta_0 + \beta_1 \,\text{Plaus} + \beta_2 \,\text{Clozedist} + \epsilon \qquad (2)$$

TABLE 8 Task performance on the comprehension questions in the EEG experiment.

	Accuracy			Reaction time		
Condition	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
A	95.1%	7.3%	75.0–100.0%	2144 ms	309 ms	1618–2781 ms
B	98.1%	6.1%	75.0–100.0%	2153 ms	316 ms	1459–3077 ms
C	95.5%	8.3%	62.5–100.0%	2182 ms	325 ms	1522–2841 ms

Note: Accuracy and reaction times were computed across subjects.

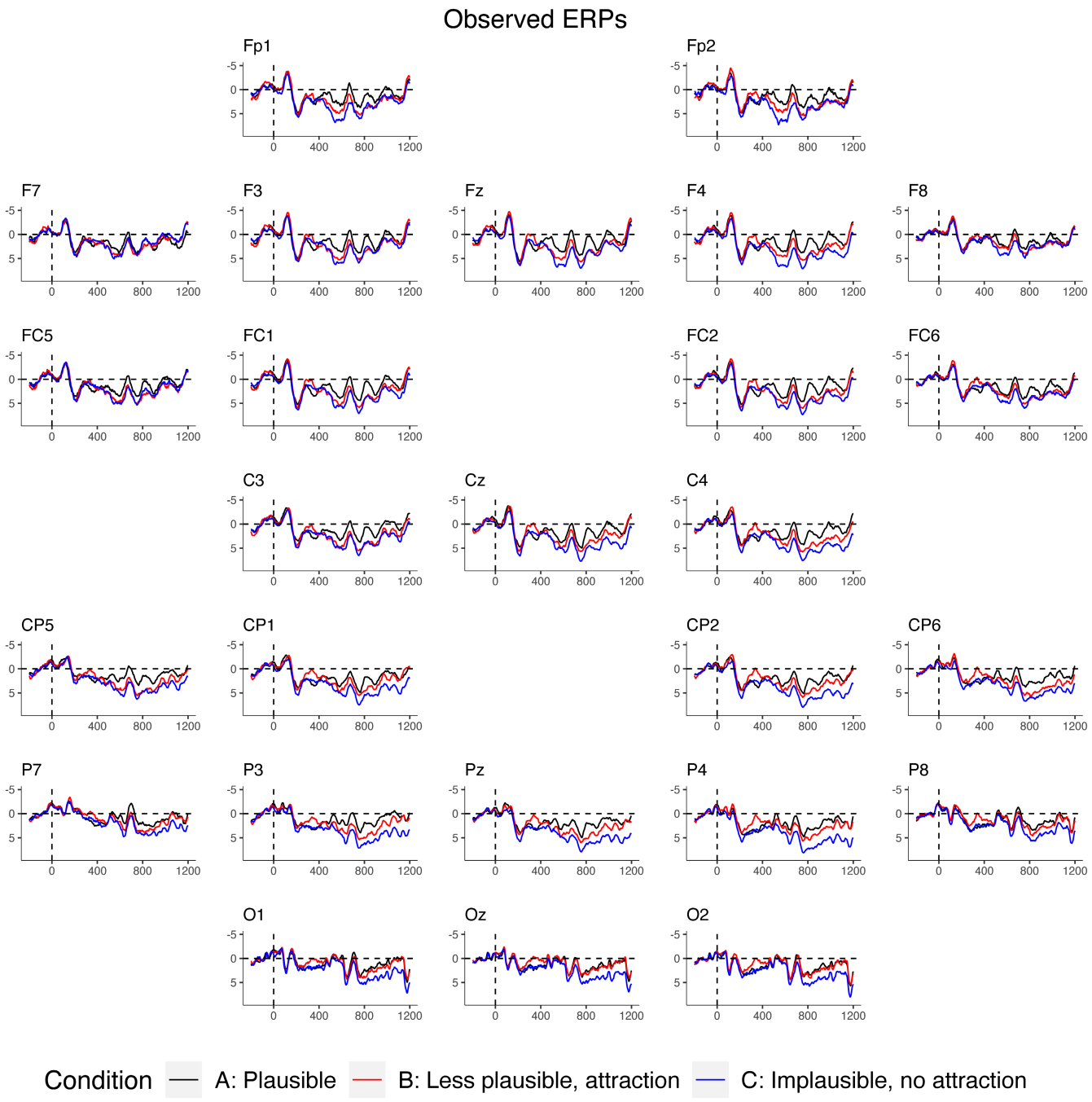

FIGURE 8 Grand-average ERPs in the three conditions manipulating plausibility and semantic attraction. Negative voltages are plotted upwards.

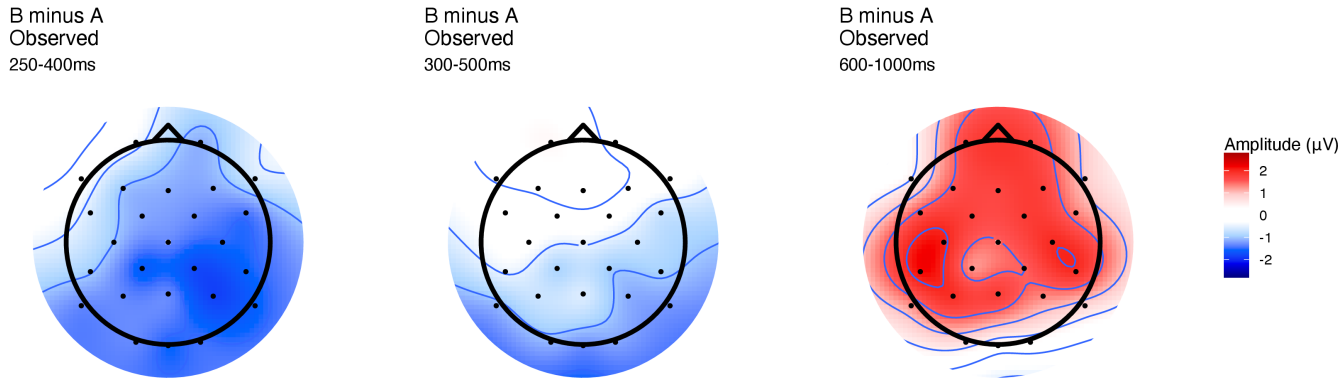

FIGURE 9 Topographic distributions of the average potentials of Condition B for the earlier negativity (250–400 ms), the canonical N400 (300–500 ms), and P600 (600–1000 ms) time windows, relative to the baseline condition. Topographies are computed from all non-reference and non-eye electrodes.

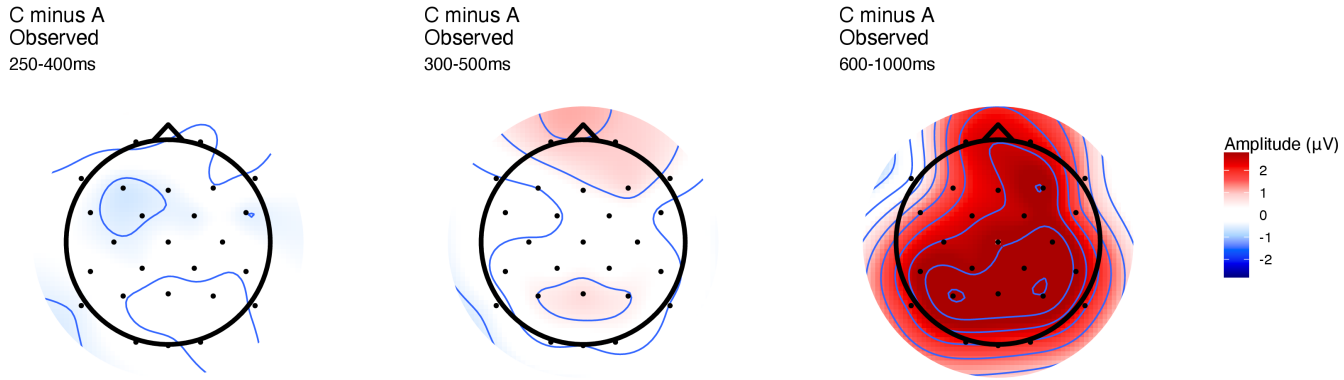

FIGURE 10 Topographic distributions of the average potentials of Condition C for the earlier negativity (250–400 ms), the canonical N400 (300–500 ms), and P600 (600–1000 ms) time windows, relative to the baseline condition. Topographies are computed from all non-reference and non-eye electrodes.

We report coefficients (β terms), estimates (the forward solution $\hat{Y}$), and residual error (ϵ, the difference between observed data Y and $\hat{Y}$), averaged across subjects. Additionally, we will compute the same models across subjects. This has the advantage that we obtain a single t value and p value for each electrode and time sample, rather than vectors of t values and p values (one value for each subject). As this still yields a multiple comparisons problem due to the multitude of time samples and electrodes, we correct p values for the inflated false discovery rate using the method proposed by Benjamini and Hochberg (1995). We adjust p values separately for the two time windows of interest but across all 26 non-reference and non-eye electrodes (Figure 8) and the time samples within a time window (N400: 300–500 ms; P600: 600–1000 ms).

3.2 | Results

3.2.1 | Comprehension questions

Participants answered comprehension questions on half of the experimental items. Descriptive metrics for accuracy and reaction times were computed across subjects. Average accuracy was 96.2% ($SD = 3.9$, range = 87.0–100.0%). Mean reaction time was 2162 ms ($SD = 254$, range = 1568–2841 ms). Accuracies and reaction times split per condition are given in Table 8.

3.2.2 | ERPs

Grand-averaged ERPs for the three conditions on all non-reference and non-eye electrodes are displayed in Figure 8. Visual inspection suggests a broadly distributed negativity, lasting approximately from 250 ms to 400 ms post-stimulus onset in response to target words that are less plausible and for which a semantically attractive alternative interpretation is present (Condition B). A smaller, more frontally pronounced early negativity, lasting approximately from 250 to 400 ms post-stimulus onset, is also evoked by implausible target words (Condition C) on frontal and central electrodes. Around the typical peak of the N400 component, no pattern of N400 amplitude with plausibility is observable by visual inspection.

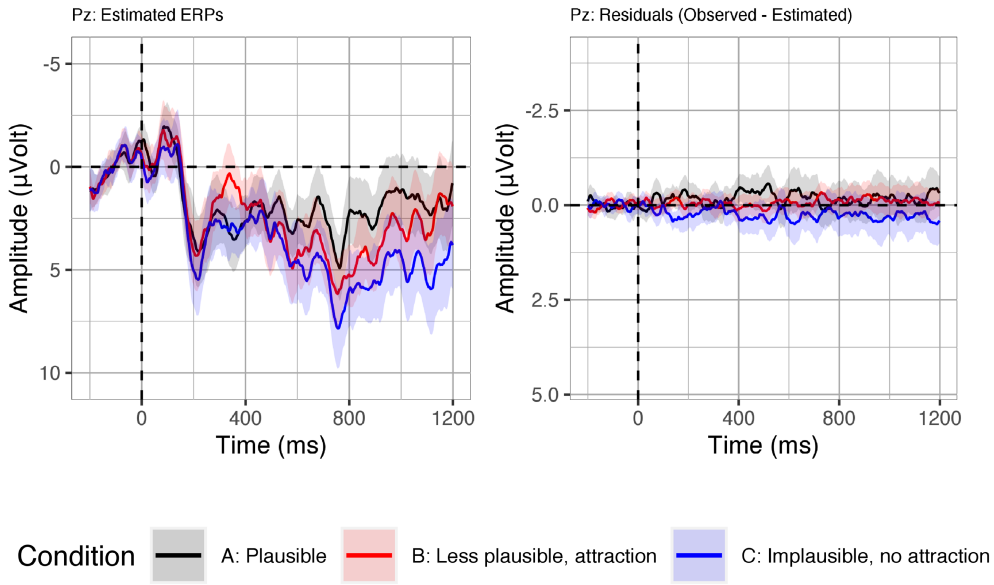

FIGURE 11 Estimated waveforms (left) and residual error (right) on electrode Pz from regression models using target word plausibility and distractor cloze probability as predictors.

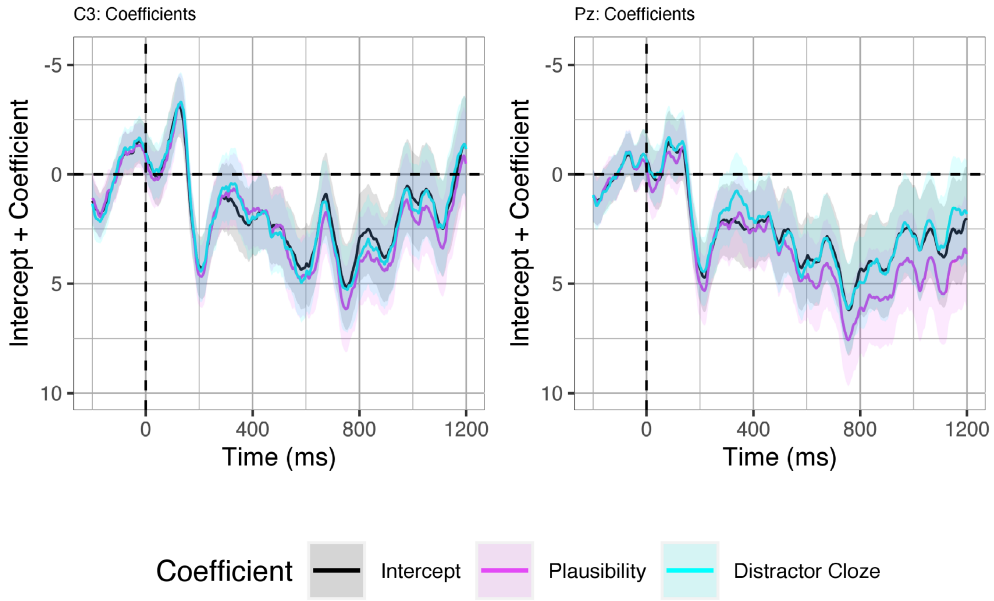

FIGURE 12 Regression model coefficients (added to their intercept) across time on electrode C3 and Pz.

Furthermore, both Condition B (less plausible, semantic attraction) and Condition C (implausible, no semantic attraction) elicit broadly distributed positivities, emerging from 500 ms post-stimulus onset. The positivity elicited by Condition C is stronger in amplitude than that elicited by Condition B on parietal electrodes. On left frontal electrodes, however, their amplitudes appear similar in parts of the epoch.

To further examine the topographies of the condition contrasts, we display topographic maps of the differences between the conditions in a time window matching visual inspection of the negativities (250–400 ms) and in the canonical N400 (300–500 ms) and P600 time windows (600–1000 ms). The topographic maps of Condition B (less plausible; semantic attraction) relative to Condition A are presented in Figure 9. The early negativity is broadly distributed and peaks over right parietal electrodes, whereas left frontally, the difference is smaller. The temporal average of the N400 time window exhibits negativities over right parietal and occipital electrodes. Inspection of the waveforms (Figure 8) strongly suggests that this negativity is driven by the temporally overlapping preceding

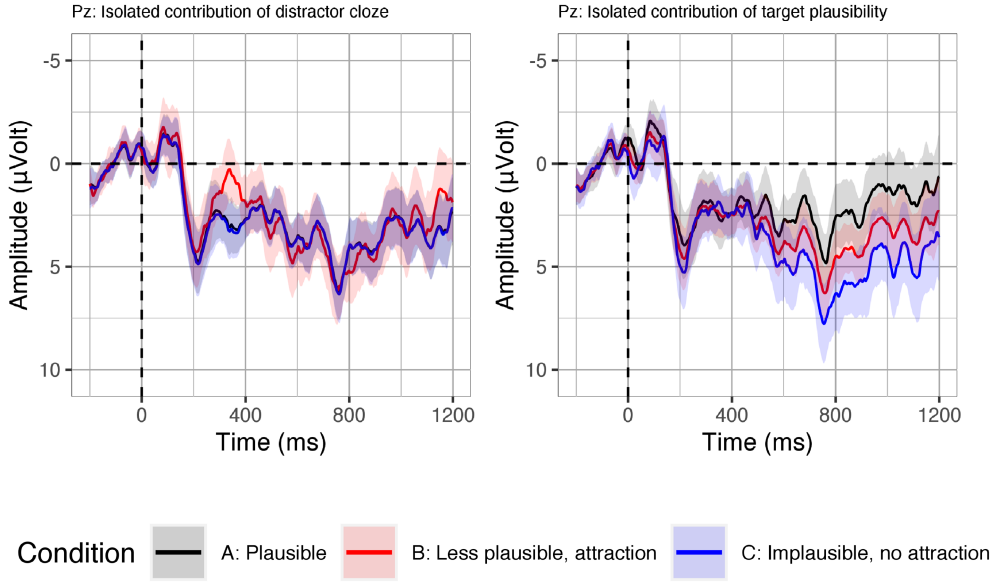

FIGURE 13 The isolated forward estimates of distractor cloze probability (left) and plausibility (right), derived from coefficients that were fitted in models containing both predictors.

negativity and that, additionally, the N400 time window also includes the onset of the P600 effect of Condition B relative to A. The late positivity has peaks both over left and right central electrodes with a trough between them.

In the topographic maps for Condition C (Figure 10), the early negativity appears much smaller than that of Condition B and peaks over left frontal electrodes. The topography in the N400 time window does not contain the topography of a typical, centrally peaking N400, but more likely shows the early, emerging P600 effect. The late positivity clearly peaks over parietal electrodes.

Turning to the rERP analysis, we first inspect the estimated waveforms for a single electrode, Pz (Figure 11; left) as well as the residual error (right), that is, the difference between the observed and the estimated data. The estimates were generated by a model with target word plausibility and distractor cloze probability as predictors. The estimates and residuals suggest that the models accurately capture the major trends in the data, as observable by visual inspection. That is, the models predict a negativity for Condition B between 250 and 400 ms, no negativity for Condition C (on this electrode), and late positivities of increasing amplitudes for Conditions B and C, respectively.

To assess which predictor captures the voltage deflections, we turn to the model coefficients, plotted over time (Figure 12; right). The coefficient for distractor cloze probability predicts the negativity elicited by Condition B, in which the distractor word was expected. Plausibility, which is stepped across the three conditions, captures the graded late positivities. In order to assess whether distractor cloze probability also predicts a late positivity on another electrode site, we also inspect the coefficients on electrode C3 (Figure 12; left), on which the late positivities for Conditions B and C appeared to match (Figure 8). Indeed, on this electrode, distractor cloze probability predicts additional positivity in parts of the P600 time window. On this electrode, plausibility also predicts a smaller earlier negativity.

Using these coefficients, we can now compute the ERPs estimated by a single predictor in isolation. To achieve this, we compute the forward estimates for the entire dataset while factoring out the influence of the other predictor by fixing it to its average value, which is zero for z-scored predictors. The isolated estimates of distractor cloze on electrode Pz contain the negativity of Condition B (Figure 13, left). Isolating the estimates of plausibility on electrode Pz (right) reveals no modulation in the N400 time window but the three-step modulation in the P600 time window. These estimates suggest that the negativity is elicited by the expectancy of the distractor word, and that plausibility predicts no N400 but P600 modulations.

As the single-electrode inspection of the coefficients suggests potential topographic differences between the contributions of the predictors, we visualize the estimated ERP data as topographic maps. This allows us to dissect how target word plausibility and distractor cloze probability interact in shaping the topographic map of the difference between Conditions B and A (see Figure 9). Figure 14 displays the individual contributions of distractor cloze probability (left), target plausibility (middle left), and their sum (middle right) to the estimated data for Condition B, which is similarly distributed to the observed data (right). The topographic maps suggest that

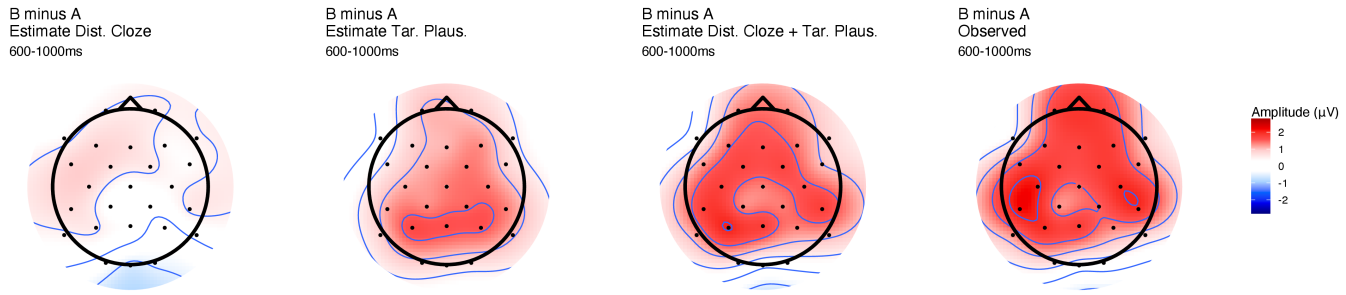

FIGURE 14 Topographic distributions of the potentials in the P600 time window estimated by distractor cloze probability (left), plausibility (middle left), and their summed estimated potentials (middle right) as well as the observed potential for Condition B (right) between 600 and 1000 ms, relative to the baseline condition. Topographies are computed from all non-reference and non-eye electrodes.

while plausibility predicts a larger, parietally peaking positivity, there is an additional left frontally peaking positivity, predicted by distractor cloze probability. This suggests that the overall topographic distribution observed for Condition B (Figure 9) is composed of a parietal and a left frontocentral subcomponent.

To assess the statistical significance of our two predictors, we computed models in which we determine the regression coefficients across all subjects, rather than fitting individual models per-subject. We report the *t* values for the two predictors on nine central electrodes (Figure 15). Furthermore, the bar below the *t* values indicates time samples that were significant after correcting for multiple comparisons within the N400 and the P600 time window and across electrodes and time samples. Our inferential statistics indicate that distractor cloze probability significantly predicts a negativity in the 300–400 ms range. While the *t* values for plausibility are large on frontal electrodes in the pre-N400 time window, indicative of a negativity predicted by low plausibility items, this does not reach significance in the current selection of time windows and electrodes. Plausibility significantly predicts a late positivity (600–1000 ms) with a peak over parietal electrodes. Distractor cloze probability, while generating a left frontocentral late positivity in the forward estimates (Figure 14), does not reach significance in our late time window.

3.3 | Discussion

Experiment 2 replicated the main findings of Nieuwland and van Berkum (2005) using visual rather than auditory language comprehension and employing an explicit task that incentivizes reading for comprehension. In the original design, a context paragraph repeatedly mentioned the target words before those same words were presented either as plausible or implausible continuations. Rather than eliciting an N400 effect, a P600 effect relative to baseline was observed. This matches our data in the less plausible condition (B: "Then *weighed* the lady the tourist") compared to the baseline (A: "Then *dismissed* the lady the tourist"). Furthermore, while a semantically attractive alternative interpretation is globally available in Condition B, it is unavailable in Condition C (C: "Then *signed* the lady the tourist"). Indeed, Condition C thus instantiates a classic semantic incongruency (see Van Petten & Luka, 2012, for a review). On multi-stream models, the absence of such a semantic attraction (Condition C) should result in the emergence of an N400 effect compared to the baseline condition. However, no N400 effect but only a P600 effect was observed in Condition C relative to A. Furthermore, our design manipulated plausibility on three levels (A: plausible < B: less plausible < C: implausible), showing that target words with intermediate plausibility ratings (B: "Then *weighed* the lady the tourist") also elicit a P600 effect, intermediate in amplitude, compared to the fully plausible and implausible conditions. Indeed, the plausibility ratings collected in a pre-test provided a continuous predictor which significantly predicted the P600 modulations observed across nine electrodes.

While distractor *absence* did not elicit an N400 effect relative to baseline, the *presence* of a distractor in fact elicited an earlier negativity, emerging from around 250 ms and lasting until 400 ms post-stimulus onset for Condition B. An interpretation of this earlier negativity as an N400 appears implausible given the temporal invariability of the N400 peak latency (Federmeier & Laszlo, 2009). Rather, we interpret this component to be elicited by the strong and unfulfilled expectation of the distractor word on a lexical level. Likely, this early component often overlaps with the N400 and it is the combination of lexical repetition and disconfirmation in our experiment that allows us to observe it in isolation. That is, even though the distractor word was strongly expected and not presented, lexical retrieval—indexed by the N400—of the target word's meaning was still maximally facilitated. Interestingly, Nieuwland and van Berkum (2005) did not observe a similar negativity in their study, even though they relied on auditory

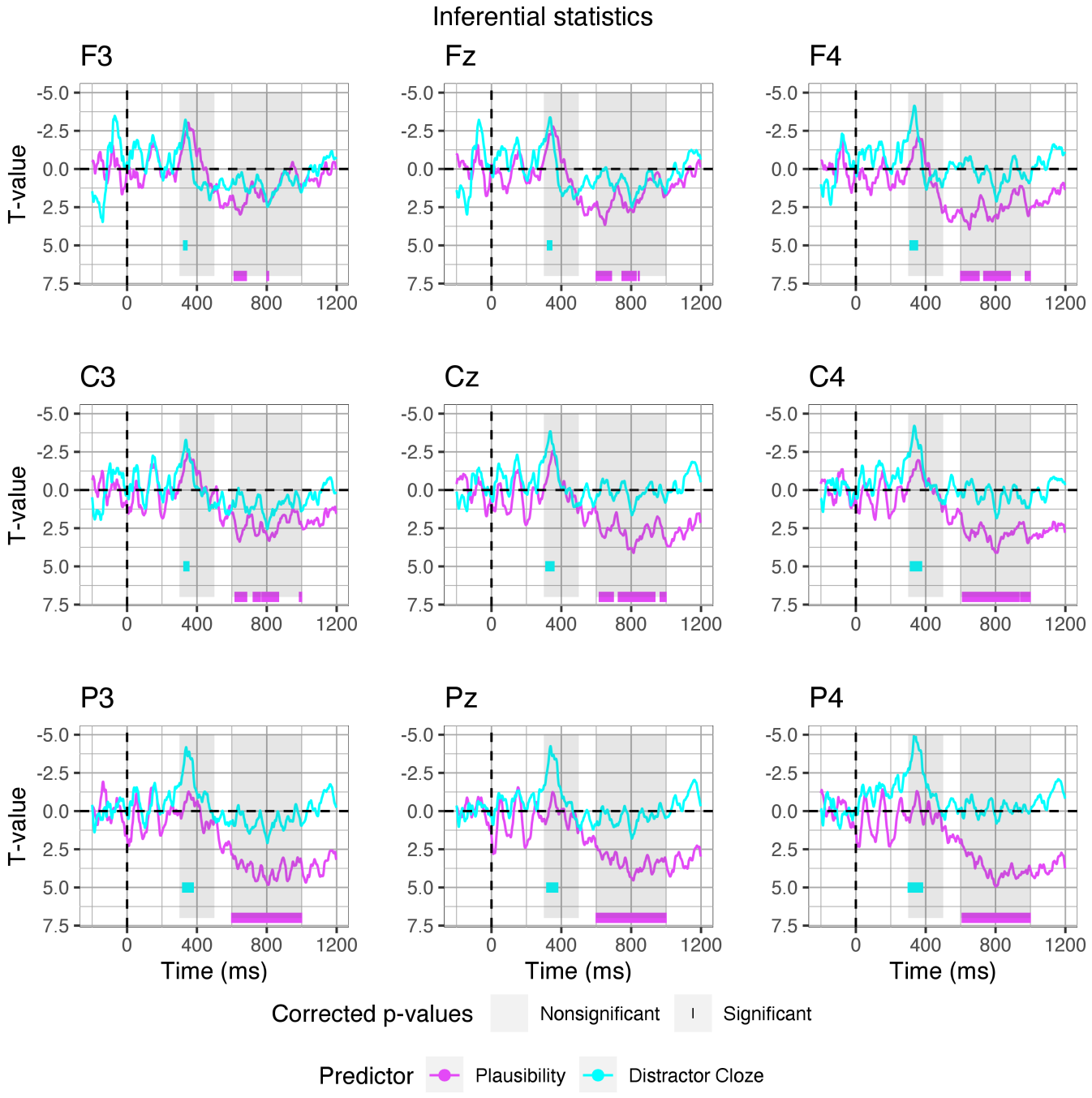

FIGURE 15 *T* values for the plausibility and distractor cloze probability predictors on nine central electrodes from across-subjects regression. Bars indicate time samples with significant *p* values after multiple comparisons correction.

presentation—a modality in which a component with a similar time course, the phonological mismatch negativity (PMN), is often observed (Connolly et al., 1990; Hagoort & Brown, 2000; Jachmann et al., 2019).

Furthermore, our rERP analyses suggest that the presence of a strongly anticipated distractor word that is then *not* presented as target word (Condition B) leads to additional modulation in the late ERP signal with a positive left frontal peak. While distractor cloze probability was not significant in the later time window, a frontal positivity could in fact be expected for our design, as the way in which our design makes a semantically attractive alternative interpretation available effectively creates a prediction disconfirmation ("Then *weighed* the lady the tourist" where "suitcase" is expected), which has been linked to frontal positivities in previous research (Brothers et al., 2015; DeLong et al., 2011, 2014; Federmeier et al., 2007; Kuperberg et al., 2020; Quante et al., 2018; see also earlier results by Kutas, 1993). Our rERP analysis suggests that the positivity observed for

Condition B can be dissected into two subcomponents: A P600 with parietal peak, predicted by plausibility, and a disconfirmation-related positivity with left-central peak, predicted by distractor cloze probability. In the design of Nieuwland and van Berkum (2005), a disconfirmation was also present; however, the replacement word was implausible. Their difference waves suggest no apparent deviation from a canonical, parietal P600. This is in line with the finding that the frontal positivity is produced by unexpected but *plausible* target words, whereas unexpected and implausible target words lead to a parietally distributed late positivity (Van Petten & Luka, 2012).

4 | GENERAL DISCUSSION

The goal of the present study was to test competing hypotheses about the functional interpretation of the N400 and P600 components. In particular, building on a previous study (Nieuwland & van Berkum, 2005), we tested the prediction of RI theory that the P600 is a continuous index of integration effort (Brouwer et al., 2017, 2021) directly against the predictions made by multi-stream models (Bornkessel-Schlesewsky & Schlesewsky, 2008; Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Michalon & Baggio, 2019; van Herten et al., 2005; and similarly Li & Ettinger, 2023; Rabovsky et al., 2018; Ryskin et al., 2021).

Multi-stream models maintain that the N400 indexes aspects of integrative/combinatorial processing of the input word with the prior context. On multi-stream accounts, no N400 modulation is generated if the processor initially does not detect an anomaly in the semantic stream because of the availability of a semantically attractive alternative interpretation. The anomaly is then detected by a second, algorithmic stream, and it is the mismatch between the analyses of the semantic stream and the algorithmic stream which produces an increase in P600 amplitude. On RI theory, by contrast, the N400 is taken to index lexical retrieval. Critically, in our design (see Table 2)—which employs a context manipulation, in which a semantically attractive alternative is either available or not (Condition B vs. C), and target word plausibility is varied across three levels (Condition A < B < C)—the target word is repeated several times in a preceding context paragraph. On the retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2008, 2009; van Berkum, 2009, 2010), this is predicted to maximally facilitate retrieval of target word meaning and thus minimize N400 differences across conditions. In sum, RI theory predicts no N400 differences across conditions, and increasing P600 amplitudes as a function of decreasing target word plausibility. Multi-stream models predict a P600 effect, but no N400 effect, if a semantically attractive alternative interpretation is available (Condition B relative to A) and an N400 effect, but no P600 effect, if no alternative interpretation is available (Condition C relative to A).

We validated the design in a self-paced reading experiment (Experiment 1) that revealed a graded sensitivity of reading times to plausibility, indicating that the stimuli indeed induce graded integration effort. Distractor cloze probability did not modulate reading speed significantly. The EEG experiment (Experiment 2), replicated the original findings of Nieuwland and van Berkum (2005), that is, the absence of an N400 effect and the presence of a P600 effect for less plausible relative to plausible target words when the target word is strongly primed by the context and in the presence of a semantically attractive alternative interpretation (our Condition B). Furthermore, our results revealed the *graded* sensitivity of a posterior late positivity to plausibility, as shown by stepped P600 amplitudes for plausible (A), less plausible (B), and implausible (C) items. The absence of a plausibility-related N400 effect is inconsistent with an interpretation of the N400 as a graded index of integration difficulty. Additionally, the presence of an expected word which was then not presented elicited an early negativity (250–400 ms)—likely a correlate of lexical mismatch. Furthermore, an rERP analysis revealed that the presence of a strongly expected distractor word—or rather its disconfirmation—resulted in an additional left-frontal positivity in a later time window, in line with previous research. However, in our analyses, the contribution of disconfirmations to late positivities was not statistically significant. In sum, as we discuss in more detail below, these findings reveal a critical novel dimension to the functional interpretation of the P600 that has important implications for existing and future neurocognitive experiments and theories, namely that the P600 is a *continuous* index of integration effort.

4.1 | The processing cost of disconfirmed expectations

While the main goal of our design was to manipulate the availability of a semantically attractive alternative interpretation (The lady weighing the suitcase rather than the tourist), the way in which we achieved this manipulation effectively created a prediction disconfirmation in Condition B. That is, when presenting the final sentence fragment "Then weighed the lady the …", "suitcase" was expected—as shown by high distractor cloze probability—but "tourist" was encountered instead. While not the main focus of our hypotheses, the results are relevant to the literature on disconfirmed predictions.

For Condition B, we observed an early negativity relative to both Condition A and C, lasting approximately from 250 to 400 ms post-stimulus onset. This deflection may relate to the mismatch between the observed word form (target) and the anticipated word form (distractor). Critically, under the retrieval view on the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2008, 2009; van Berkum, 2009, 2010), this mismatch does not appear to tax lexical retrieval, as no N400 modulation was observed: The difference between the waveforms disappeared by 400 ms, which would be the typical peak of the N400 component (Federmeier & Laszlo, 2009). This earlier negative component likely overlaps with the N400 in previous studies on disconfirmations and it is the absence of an N400 effect relative to baseline in our data that allows us to observe the earlier negativity in isolation. Results that are directly relevant to ours are presented by Brothers et al. (2015), who observed a centrally peaking N250 for the contrast between a medium-cloze unpredicted versus a medium-cloze predicted target word. Furthermore, in their data, the earlier negativity was not observed for the contrast of a low-cloze unpredicted to a medium-cloze unpredicted target word, which only elicited an N400 effect. Similarly, the visual mismatch negativity has been reported for exactly the time window between 250 ms and 400 ms (Tales et al., 1999). Furthermore, negativities preceding the N400 time window have been found for expectation-incompatible relative to expectation-compatible stimuli (Bartholow et al., 2005), for expectation-based semantic priming (Franklin et al., 2007), and, using pictorial stimuli, for perceptual hypothesis testing which is argued to precede multimodal semantic memory access, as indexed by the following N400 (Kumar et al., 2021).

In the time window from 600 to 1000 ms, our rERP analysis suggests that target words that disconfirmed expected distractor words induced a left frontal positivity. Distractor cloze probability did, however, not reach significance in the analyses, and hence, these results warrant adequate caution. Nevertheless, previous research has repeatedly reported frontal positivities elicited by prediction disconfirmations (Brothers et al., 2015; DeLong et al., 2011, 2014; Federmeier et al., 2007; Kuperberg et al., 2020; Kutas, 1993; Quante et al., 2018), making our results relevant to this line of research. A prominent idea has been that if the target is unexpected but plausible, disconfirmations result in a frontally pronounced positivity, whereas implausible replacements result in a parietal positivity (Van Petten & Luka, 2012). We see, however, two open issues with regard to this strict functional segregation of frontal and parietal positivities. First, the apparent distinction between frontally and parietally distributed positivities could be an artifact of spatiotemporal component overlap with the N400 (Brouwer & Crocker, 2017; Delogu et al., 2021), and second, frontally and parietally distributed positivities may not be mutually exclusive.

A relevant study by DeLong et al. (2014) included plausible, less plausible disconfirming, and implausible disconfirming target words. The design elicited a frontal positivity for less plausible disconfirming words, a parietal positivity for implausible disconfirming words, and, critically, N400 effects in response to both less plausible and implausible words, relative to baseline. Our design does not elicit N400 differences and hence circumvents the issue of component overlap, thereby providing a clearer view on the distribution of the late positivities. The estimates generated by our rERP models (Figure 14) suggest that even without a strong N400 overlapping with the late positivity, unfulfilled expectations create an additional positivity with a left-frontocentral distribution. Furthermore, in the disconfirming condition (B), the context additionally made the target word less plausible compared to the baseline condition. Our rERP analysis revealed that for Condition B, plausibility induces a parietal P600—which was not observed in the data of DeLong et al. (2014)—in addition to the frontal positivity elicited by the disconfirmation. In sum, our results and the rERP analysis suggest that disconfirmations indeed induce a frontal positivity, but that this frontal positivity can co-occur with a plausibility-related parietal positivity on less plausible, but ultimately possible target words.

4.2 | Global revision on the multi-stream account

The main goal of this study was to test the hypotheses of multi-stream models against those of RI theory. Multi-stream models were originally proposed in response to studies eliciting semantic P600s, in which semantic anomalies did not elicit N400 effects but rather P600 effects, relative to baseline. Multi-stream accounts explain some of the original data points, by postulating that the semantic stream does not detect the anomaly because a semantically attractive alternative interpretation is available. For instance, in order to "repair" the sentence "the hearty meal was devouring," the inflection of the verb could be changed to "devoured," yielding a plausible interpretation. However, the surface structure of the sentence does not match this interpretation, which is detected by the algorithmic stream and the conflict between the two streams leads to a P600 effect when compared to a congruous condition.

This explanation was based on a *locally* available alternative interpretation (see Figure 2). However, no such *local* availability is given in the design of Nieuwland and van Berkum (2005, "Next, the lady told

the tourist/suitcase), and, accordingly, an N400 and no P600 effect relative to baseline would be predicted by multi-stream models. However, the reverse pattern was observed. To account for this, multi-stream may invoke a *globally* attractive alternative interpretation (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Kuperberg, 2007, for discussion). That is, making use of the *globally* available information, the word "suitcase" could be replaced with the discourse-salient word "tourist" in order to arrive at a plausible interpretation in the semantic stream. Again, the analysis generated by the algorithmic stream conflicts with the analysis of the semantic stream, explaining the P600 increase found by Nieuwland and van Berkum (2005). Importantly, it follows that if neither a *locally* nor a *globally* available alternative interpretation is present, an N400 effect should be observed relative to baseline.

The current study adapted the original design by Nieuwland and van Berkum (2005) to test this prediction. In the new context manipulation design, we made an alternative interpretation available *globally* for a less plausible target word (Condition B: "Next, the lady weighed the tourist"), whereas no alternative interpretation was available for the fully implausible target word (Condition C: "Next, the lady signed the tourist"). Assuming a plausibility heuristic aware of globally available alternatives, multi-stream models predict only a P600 effect for Condition B and only an N400 effect for Condition C relative to Condition A. Note that multi-stream models in general predict either an N400 or a P600 increase, which makes biphasic N400-P600 results problematic for most multi-stream accounts (see Van Petten & Luka, 2012, for an overview, Brouwer et al., 2012, for discussion, and Bornkessel-Schlesewsky & Schlesewsky, 2008; Li & Ettinger, 2023, for exceptions).

In Condition B, for which only a P600 is predicted by multi-stream accounts, we found a P600 effect relative to Condition A. This condition replicates the results of Nieuwland and van Berkum (2005), and, accordingly, multi-stream models can only explain this P600 effect by invoking a *globally* available alternative interpretation. In Condition C, for which only an N400 effect is predicted by multi-stream accounts, we observed only a P600 effect, relative to Condition A. Critically, the absence of an N400 effect relative to baseline when any semantically attractive alternative interpretation is removed provides strong evidence against multi-stream accounts. One explanation of the absence of the N400 effect in Condition C relative to A would be to assume that the revision process changed the context of Condition C ("Then *signed* the lady the") to make the target word ("tourist") plausible. It is difficult, however, to imagine a mechanism that could revise the context in such a way, while at the same time predicting the presence of N400 effects in cases of canonical semantic incongruencies (see Van Petten & Luka, 2012). Another explanation would entail misunderstanding "tourist" for something contextually relevant, such as the "tourist's ticket". Many of our stimuli, however, contain strong selectional restriction violations, such as "the apprentice ate the hammer" (see Appendix S1), where reference transfer to a thus far unnamed entity seems unlikely, and hence, this explanation cannot account for the complete absence of an N400 effect of Condition C relative to A. Again, it is difficult to see how such an account would predict the absence of an N400 effect for the present stimuli, while at the same predicting the presence of an N400 effect for canonical semantic incongruencies. In sum, we do not see how the present data can be reconciled with the mechanisms assumed by multi-stream accounts.

4.3 | Retrieval facilitation under repetition priming

The current design had the goal of maximally priming the target word by mentioning it repeatedly in a context paragraph preceding the final sentence. The prediction of RI theory was that maximal priming should maximally facilitate retrieval of the target word's meaning from long-term memory, thus leading to equal N400 amplitudes across conditions. Our results revealed that while an earlier negativity was present in Condition B relative to A (see above), no difference in the canonical N400 time window was observed for any condition contrast—in line with the retrieval view of the N400 (Brouwer et al., 2012; Kutas & Federmeier, 2000, 2011; Lau et al., 2008, 2009; van Berkum, 2009, 2010). This study thus adds to several studies that elicited no N400 differences for target words that were equally strongly or weakly primed by the preceding context (Delogu et al., 2019, 2021; Hoeks et al., 2004; Kim & Osterhout, 2005; Kos et al., 2010; Kuperberg, 2007; Nieuwland & van Berkum, 2005; Otten & van Berkum, 2008; van Herten et al., 2005).

Critically, our results show that even when the target word is of intermediate plausibility (Condition B) or entirely implausible (Condition C), no N400 increase is produced—a result that is at odds with the traditional interpretation of the N400 as semantic integration (Brown & Hagoort, 1993, 2000; Hagoort et al., 2004). Furthermore, also when assuming a hybrid view of the N400 that takes the N400 to index both retrieval and aspects of integrative processing (see Baggio & Hagoort, 2011, who refer to this as "unification," and Baggio, 2018, for an updated account), we would expect to find N400 modulations for the less plausible or implausible target words even when their word meaning

is strongly and equally primed—a prediction which was not confirmed. That is, even though retrieval may be facilitated, these accounts should still predict increased integration effort to be reflected in the N400. Thus, for hybrid models to predict the absence of any N400 effect of implausibility, they must still assume that retrieval processes dominate integration/unification. While it may be possible to construct such a hybrid account, the data are more parsimoniously explained by a retrieval only account, and we are unaware of any other findings that necessitate the inclusion of an integration mechanism. Moreover, it is difficult to see how such an account can explain the absence of an N400 effect of implausibility, when target words are equally unassociated with the context (Delogu et al., 2021). Another proposal by Nieuwland et al. (2020) suggests that the earlier part of the N400 is sensitive to retrieval processes, while the later part indexes integration. Critically, however, we did not observe any N400 differences in either the earlier or later part of this component, thereby also ruling out this proposition. On a final note, the absence of N400 modulations by plausibility supports the view that the correlation between corpus-based word surprisal and the N400 may be best explained by expectation-based modulations of lexical retrieval rather than integration (see Aurnhammer et al., 2021; Frank et al., 2015, for discussion).

4.4 | The P600 as a graded index of integration effort

Most strikingly, our ERP data revealed an important novel dimension of the P600 component: Our design manipulated plausibility on three levels (plausible, less plausible, implausible) and revealed that P600 amplitude patterns with plausibility. Going beyond the three discrete levels of plausibility, we successfully modeled the ERP signal as a continuous function of numeric per-item plausibility ratings collected in a pre-test, indicating that the P600 may indeed be a continuous index of integration effort. We conclude that P600s are not only elicited by highly implausible, impossible, or violating target words (Bornkessel-Schlesewsky et al., 2011; Kuperberg, 2007), but rather, that P600 amplitude is modulated as a function of integration effort by every word.

Our proposition that the P600 is a continuous index of integration effort is indeed supported by numerous previous studies showing P600 effects for non-violating but semantically or pragmatically taxing continuations (Burkhardt, 2006, 2007; Cohn & Kutas, 2015; Delogu et al., 2019; Dimitrova et al., 2012; Hoeks et al., 2013; Regel et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Xu & Zhou, 2016). For instance, a world knowledge implausibility without a violation of selectional restrictions induced a P600 effect relative to control (Delogu et al., 2019). The graded nature of the P600 was also suggested by a post hoc analysis conducted by Aurnhammer et al. (2021). By analyzing the data of the baseline condition only ("Yesterday sharpened the lumberjack [...] the axe", translated from German), it was found that not only the N400 but also the P600 varied gradually as a function of target word expectancy. This observation was interpreted as indicating a gradual modulation of lexical retrieval (N400) and integration (P600) by expectancy. Hence, the current study directly supports their exploratory, post hoc analysis with regard to the P600 component.

In Experiment 1, the observed reading times closely patterned with the P600s in that both were modulated by plausibility on the three levels of our manipulation. Taken together with the absence of N400 modulations by plausibility, this strengthens the proposed link between reading times and the P600 through comprehension-centric surprisal (Brouwer et al., 2021). To further test this idea, we conduct a post hoc analysis, in which we apply the rERP technique to model the ERPs obtained in Experiment 2 by the reading times obtained on the Post-spillover region in Experiment 1 (averaged per item and condition). The resulting coefficients (Figure 16) suggest that indeed, the observed positivities are correlated to the observed reading times, suggesting they may be closely associated indices of

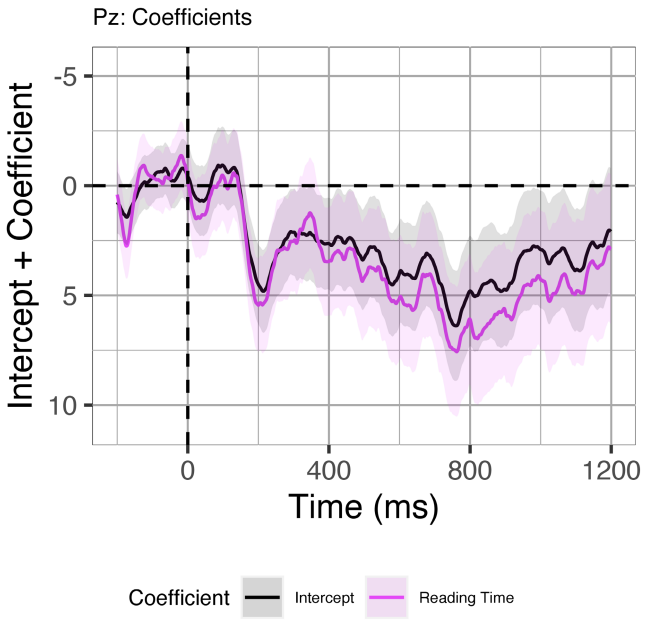

FIGURE 16 Regression model coefficients (added to their intercept) across time on electrode Pz from models predicting the ERPs as a function of the per-item and per-condition reading times obtained on the Post-spillover region in Experiment 1.

processing effort across pools of participants. This finding further corroborates the P600 as a continuous index of integration effort.

5 | CONCLUSION

Event-related potentials provide a multidimensional window into the nature and time course of language comprehension. Critically, establishing the locus of specific subprocesses of comprehension in the ERP signal has direct consequences for our understanding of the temporal organization and architecture of the comprehension system. The present study directly tested competing views on whether the N400 or the P600 component of the ERP signal indexes the integration of incoming word meaning into an unfolding utterance representation. Crucially, the traditional view of the N400 as an index of integration relies on the presence of a semantically attractive alternative interpretation to explain the absence of an N400 effect in response to certain semantic anomalies. The more recent view of the P600 as an index of integration, in turn, predicts P600 amplitude to be a continuous index of integration effort, a prediction that had yet to be confirmed. We harnessed these predictions to decide between the competing views using a design in which a semantically attractive alternative is either available or not, and target word plausibility is varied across three levels. Furthermore, to minimize lexical processing differences across conditions, target words were equally primed by the prior context.

An initial self-paced reading study revealed a gradual slowdown of reading times for gradual decreases in target word plausibility, suggesting differential integration effort. In the ERP study, the plausibility manipulation did not elicit any N400 differences across conditions. Indeed, the lack of an increased N400 for the implausible conditions—even when no locally or globally attractive alternative interpretation is available—is directly at odds with the prediction made by contemporary models that maintain the N400 as an index of semantics-driven, "quasi-compositional" integration. In fact, the plausibility manipulation rather revealed P600 amplitude to be graded for plausibility. Taken together, these results cannot be reconciled with the N400 as an index of integration, while they are consistent with the P600 as a continuous index of integrative effort. More generally, the results are consistent with Retrieval–Integration theory, a single-stream account in which the N400 indexes lexical retrieval from long-term memory and the P600 indexes integration of incoming word meaning into an unfolding utterance representation. No N400 differences were found, as lexical retrieval was equally facilitated across conditions through repetition priming, and the link between plausibility, reading times, and P600 amplitude establishes the P600 as a direct index of semantic integration that—in line with a comprehension-centric notion of surprisal—is continuous in amplitude as a function of integration effort. This novel dimension of the P600 has important implications for existing and future experiments, as well as for theories and models of language comprehension.

AUTHOR CONTRIBUTIONS

Christoph Aurnhammer: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Francesca Delogu:** Data curation; formal analysis; methodology; resources; validation; writing – review and editing. **Harm Brouwer:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; software; supervision; validation; writing – review and editing. **Matthew W. Crocker:** Conceptualization; funding acquisition; methodology; project administration; supervision; writing – review and editing.

ACKNOWLEDGMENTS

We would like to thank Mante Nieuwland for providing the materials of the Nieuwland and van Berkum (2005) study. We also thank Lea Müller-Kirchen for her help in designing the materials and conducting the ERP study. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

DATA AVAILABILITY STATEMENT

Code and data required to reproduce the analyses are publicly available at https://github.com/caurnhammer/psyp23rerps.

ORCID

Christoph Aurnhammer https://orcid.org/0000-0003-1898-6253
Francesca Delogu https://orcid.org/0000-0002-8158-126X
Harm Brouwer https://orcid.org/0000-0002-7336-4142
Matthew Crocker https://orcid.org/0000-0003-3452-3064

REFERENCES

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLoS One*, *16*(9), e0257430. https://doi.org/10.1371/journal.pone.0257430

Baggio, G. (2018). *Meaning in the brain*. MIT Press.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338–1367. https://doi.org/10.1080/01690965.2010.542671

Bartholow, B. D., Pearson, M. A., Dickter, C. L., Sher, K. J., Fabiani, M., & Gratton, G. (2005). Strategic control and medial frontal negativity: Beyond errors and response conflict. *Psychophysiology*, *42*(1), 33–42. https://doi.org/10.1111/j.1469-8986.2005.00258.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., Roehm, D., & Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152. https://doi.org/10.1016/j.bandl.2010.09.010

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. https://doi.org/10.1016/j.brainresrev.2008.05.003

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

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, *8*, 1327. https://doi.org/10.3389/fpsyg.2017.01327

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*, 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, *12*, 615538. https://doi.org/10.3389/fpsyg.2021.615538

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34

Brown, C., & Hagoort, P. (2000). On the electrophysiology of language comprehension: Implications for the human language system. In M. W. Crocker, M. Pickering, & C. J. Clifton (Eds.), *Architectures and mechanisms for language processing* (pp. 213–237). Cambridge University Press.

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159–168. https://doi.org/10.1016/j.bandl.2006.04.005

Burkhardt, P. (2007). The P600 reflects cost of new information in discourse memory. *Neuroreport*, *18*(17), 1851–1854. https://doi.org/10.1097/WNR.0b013e3282f1a999

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia*, *77*, 267–278. https://doi.org/10.1016/j.neuropsychologia.2015.08.026

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language*, *39*(2), 302–318. https://doi.org/10.1016/0093-934X(90)90016-A

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition*, *135*, 103569. https://doi.org/10.1016/j.bandc.2019.05.007

Delogu, F., Brouwer, H., & Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, *1766*, 147514. https://doi.org/10.1016/j.brainres.2021.147514

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., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

Dimitrova, D. V., Stowe, L. A., Redeker, G., & Hoeks, J. C. J. (2012). Less is not more: Neural responses to missing and superfluous accents in context. *Journal of Cognitive Neuroscience*, *24*(12), 2400–2418. https://doi.org/10.1162/jocn_a_00302

Eddine, S. N., Brothers, T., & Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. In K. D. Federmeier (Ed.), *Psychology of learning and motivation* (Vol. *76*, pp. 123–206). Academic Press. https://doi.org/10.1016/bs.plm.2022.03.005

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. *51*, pp. 1–44). Academic Press. https://doi.org/10.1016/S0079-7421(09)51001-8

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

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Franklin, M. S., Dien, J., Neely, J. H., Huber, E., & Waterson, L. D. (2007). Semantic priming modulates the N400, N300, and N400RP. *Clinical Neurophysiology*, *118*(5), 1053–1068. https://doi.org/10.1016/j.clinph.2007.01.012

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214. https://doi.org/10.1016/j.jml.2017.04.007

Fritz, I., & Baggio, G. (2020). Meaning composition in minimal phrasal contexts: Distinct ERP effects of intensionality and

denotation. *Language, Cognition and Neuroscience*, *35*(10), 1295–1313. https://doi.org/10.1080/23273798.2020.1749678

Fritz, I., & Baggio, G. (2022). Neural and behavioural effects of typicality, denotation and composition in an adjective–noun combination task. *Language, Cognition and Neuroscience*, *37*(5), 537–559. https://doi.org/10.1080/23273798.2021.2004176

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. https://doi.org/10.1080/01690960902965951

Hagoort, P., Baggio, G., & Willems, R. M. (2009). *Semantic unification*. MIT Press.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia*, *38*(11), 1518–1530. https://doi.org/10.1016/S0028-3932(00)00052-X

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19*(1), 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022

Hoeks, J. C. J., Stowe, L. A., Hendriks, P., & Brouwer, H. (2013). Questions left unanswered: How the brain responds to missing information. *PLoS One*, *8*(10), e73594. https://doi.org/10.1371/journal.pone.0073594

Jachmann, T. K., Drenhaus, H., Staudte, M., & Crocker, M. W. (2019). Influence of speakers' gaze on situated language comprehension: Evidence from event-related potentials. *Brain and Cognition*, *135*, 103571. https://doi.org/10.1016/j.bandc.2019.05.009

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36. https://doi.org/10.1016/S0093-934X(02)00548-5

Kos, M., Vosse, T. G., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00222

Kumar, M., Federmeier, K. D., & Beck, D. M. (2021). The N300: An index for predictive coding of complex visual objects and scenes. *Cerebral Cortex Communications*, *2*(2), tgab030. https://doi.org/10.1093/texcom/tgab030

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., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., & Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, *15*(2), 272–293. https://doi.org/10.1162/089892903321208204

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., & 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*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for N400 context effects: Evidence from MEG. *Brain and Language*, *111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

Lau, E., 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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, 105359. https://doi.org/10.1016/j.cognition.2022.105359

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

Michalon, O., & Baggio, G. (2019). Meaning-driven syntactic predictions in a parallel processing architecture: Theory and algorithmic modeling of ERP effects. *Neuropsychologia*, *131*, 171–183. https://doi.org/10.1016/j.neuropsychologia.2019.05.009

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics* (pp. 398–408). Association for Computational Linguistics.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Matthew Husband, E., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., & van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*(3), 691–701. https://doi.org/10.1016/j.cogbrainres.2005.04.003

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

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

Quante, L., Bölte, J., & Zwitserlood, P. (2018). Dissociating predictability, plausibility and possibility of sentence continuations in reading: Evidence from late-positivity ERPs. *PeerJ*, *6*, e5717. https://doi.org/10.7717/peerj.5717

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

Regel, S., Gunter, T. C., & Friederici, A. D. (2010). Isn't it ironic? An electrophysiological exploration of figurative language processing. *Journal of Cognitive Neuroscience*, *23*(2), 277–293. https://doi.org/10.1162/jocn.2010.21411

Rich, S., & Harris, J. (2021). Unexpected guests: When disconfirmed predictions linger. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2246–2252). Cognitive Science Society.

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*, 107855. https://doi.org/10.1016/j.neuropsychologia.2021.107855

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Incorporated.

Schumacher, P. B. (2011). The hepatitis called ...: Electrophysiological evidence for enriched composition. In J. Meibauer & M. Steinbach (Eds.), *Experimental pragmatics/semantics* (pp. 199–2019). John Benjamins Publishing.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344–354. https://doi.org/10.1037/0278-7393.14.2.344

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

Spotorno, N., Cheylus, A., Henst, J.-B. V. D., & Noveck, I. A. (2013). What's behind a P600? Integration operations during irony processing. *PLoS One*, *8*(6), e66839. https://doi.org/10.1371/journal.pone.0066839

Tales, A., Newton, P., Troscianko, T., & Butler, S. (1999). Mismatch negativity in the visual modality. *Neuroreport*, *10*(16), 3363–3367. https://doi.org/10.1097/00001756-199911080-00020

van Berkum, J. J. A. (2009). The neuropragmatics of 'simple' utterance comprehension: An ERP review. In U. Sauerland & K. Yatsushiro (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave Macmillan.

van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad – any implications for neuropragmatics? *Italian Journal of Linguistics*, *22*, 181–208.

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 Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, *18*(7), 1181–1197. https://doi.org/10.1162/jocn.2006.18.7.1181

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., & 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

Xu, X., & Zhou, X. (2016). Topic shift impairs pronoun resolution during sentence comprehension: Evidence from event-related potentials. *Psychophysiology*, *53*(2), 129–142. https://doi.org/10.1111/psyp.12573

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1

How to cite this article: Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *00*, e014302. https://doi.org/10.1111/psyp.14302