Received: 18 August 2020 | Revised: 28 August 2020 | Accepted: 29 August 2020

DOI: 10.1111/ejn.14961

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

Splitting event-related potentials: Modeling latent components using regression-based waveform estimation

Harm Brouwer | **Francesca Delogu** | **Matthew W. Crocker**

Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

Correspondence
Harm Brouwer, Department of Language Science and Technology, Saarland University, 66123, Saarbrücken, Germany.
Email: brouwer@coli.uni-saarland.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 232722074 (SFB1102)

Abstract

Event-related potentials (ERPs) provide a multidimensional and real-time window into neurocognitive processing. The typical Waveform-based Component Structure (WCS) approach to ERPs assesses the modulation pattern of components—systematic, reoccurring voltage fluctuations reflecting specific computational operations—by looking at mean amplitude in predetermined time-windows. This WCS approach, however, often leads to inconsistent results within as well as across studies. It has been argued that at least some inconsistencies may be reconciled by considering spatiotemporal overlap between components; that is, components may overlap in both space and time, and given their additive nature, this means that the WCS may fail to accurately represent its underlying latent component structure (LCS). We employ regression-based ERP (rERP) estimation to extend traditional approaches with an additional layer of analysis, which enables the explicit modeling of the LCS underlying WCS. To demonstrate its utility, we incrementally derive an rERP analysis of a recent study on language comprehension with seemingly inconsistent WCS-derived results. Analysis of the resultant regression models allows one to derive an explanation for the WCS in terms of how relevant regression predictors combine in space and time, and crucially, how individual predictors may be mapped onto unique components in LCS, revealing how these spatiotemporally overlap in the WCS. We conclude that rERP estimation allows for investigating how scalp-recorded voltages derive from the spatiotemporal combination of experimentally manipulated factors. Moreover, when factors can be uniquely mapped onto components, rERPs may offer explanations for seemingly inconsistent ERP waveforms at the level of their underlying latent component structure.

KEYWORDS

event-related potentials (ERPs), N400, P600, regression-based ERPs (rERPs), spatiotemporal component overlap

Abbreviations: ANOVA, analysis of variance; AP, anterior–posterior; DBC, Delogu, Brouwer, and Crocker (2019); EROS, event-related optical signal; ERP, event-related potential; LCS, latent component structure; lIFG, left inferior frontal gyrus; LMER, linear mixed effects regression; lpMTG, left posterior middle temporal gyrus; MVPA, multivariate pattern analysis; rERP, regression-based event-related potential; ROI, region of interest; SE, standard error; VIF, variance inflation factor; WCS, waveform-based component structure.

[Correction added on 09 October 2020, after first online publication: Figures 3 and 8 were swapped in the original publication. This has been amended.]

1 | INTRODUCTION

Event-related potentials (ERPs) offer a high resolution, multidimensional window into the time course of neurocognitive processing. Of interest in the ERP signal are systematic, reoccurring voltage fluctuations called *components*, which are taken to reflect the neural activity underlying specific computational operations carried out in given neuroanatomical networks (cf. Luck, 2005a; Näätänen & Picton, 1987; Rugg & Coles, 1995). Indeed, many questions in cognitive science are concerned with how the modulation pattern of such components in the signal—identified by looking at mean amplitude in predetermined time-windows—differs between experimental manipulations (Handy, 2005; Luck, 2005a; Luck & Kappenman, 2012). This approach, however, often leads to inconsistent results within (e.g. Delogu, Brouwer, & Crocker, 2019; Kim & Osterhout, 2005; Kolk, Chwilla, Van Herten, & Oor, 2003; Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007) as well as as across studies (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer, Fitz, & Hoeks, 2012; Kuperberg, 2007; van Petten & Luka, 2012, for reviews).

Crucially, such inconsistencies in Waveform-based Component Structure (WCS)—component structure derived from the ERP waveforms—may be reconciled by factoring in spatiotemporal overlap between components (e.g. Brouwer & Crocker, 2017; Donchin, Ritter, & McCallum, 1978; Duncan et al., 2009; Luck, 2005a, 2005b; Näätänen, 1982; Squires, Squires, & Hillyard, 1975). That is, computational operations may be carried out simultaneously, and even dynamically interact, which means that the ERP components that index these operations will overlap in both space and time (e.g. Hagoort, 2003; Luck, 2005a, 2005b). Given that ERP components are additive, a direct implication of this spatiotemporal overlap is that WCS may look very different from its underlying latent component structure (LCS; see Brouwer & Crocker, 2017, for discussion). While the importance of focusing on LCS is generally acknowledged, investigating LCS has proven very challenging, as the scalp-recorded ERP signal inherently conflates the contributions of different components. Indeed, traditional approaches to analyzing ERPs, for example, using ANOVAs or linear mixed models, are therefore limited to assessing WCS. Furthermore, even more complex approaches such as "decoding" using multivariate pattern analysis (MVPA; see King & Dehaene, 2014, for an introduction), which have been proposed as a viable tool to assess LCS (Heikel, Sassenhagen, & Fiebach, 2018), do not truly decompose the observed signal into its different constituent sources. Hence, analyses using such methods thus remain at best suggestive. In the present paper, by contrast, we show how traditional approaches can be extended with an additional layer of analysis that does enable true decomposition of WCS into it's underlying LCS. To this end, we employ the rERP estimation framework proposed by Smith and Kutas (2015a) (and Smith & Kutas, 2015b), a powerful extension of traditional ERP averaging that naturally allows for dealing with designs that combine categorical and continuous covariates, the correction of overlap due to temporally adjacent stimuli, effects that are non-linear, and more. We show how this rERP framework can be harnessed to explicitly model the LCS underlying WCS, and how such an LCS analysis allows one to reconcile the seemingly internally inconsistent results from a recent study on language comprehension by Delogu et al. (2019, henceforth DBC).[1]

DBC examined the effect of plausibility (the degree to which an utterance makes sense in light of our knowledge of the world) and semantic association (the degree to which a critical word is related in meaning to the prior context) on the processing of mini-discourses. A first contrast looked at the effect of manipulating plausibility only ("John [left/entered] the restaurant. Before long, he opened the menu […]"). Relative to the plausible baseline (opening the menu after entering the restaurant), the critical word "menu" in the implausible condition (opening the menu after leaving the restaurant) produced an effect on the amplitude of the P600 component, a positive deflection in the signal starting around 600 ms post word-onset. A second contrast looked at the effect of manipulating both plausibility and semantic association ("John entered the [apartment/restaurant]. Before long, he opened the menu […]"). Here, the critical word "menu" in the implausible and semantically unassociated condition (opening the menu after entering the apartment) produced an effect on the amplitude of the N400 component, a negative deflection in the signal between 300 and 500 ms, which continued into a sustained negativity, and a small positivity at occipital electrodes only. Hence, while the first contrast seems to clearly suggest that plausibility is reflected in P600 amplitude, the second contrast calls this into question. The DBC results thus appear internally inconsistent, and inconsistencies such as these are widespread, both within and across studies (Brouwer & Crocker, 2017; van Petten & Luka, 2012).

One could seek an explanation for such apparent inconsistencies in cognitive architecture. That is, in the DBC study, it could be that plausibility and semantic association interact in such a way that their combined manipulation leads to a qualitatively different type of processing, and hence a qualitatively different modulation of the ERP signal: one in which plausibility does not affect the amplitude of the P600 component, or at least not at the same electrode sites and to the same degree as when only plausibility is

[1]See 'Data Availability Statement' below.

manipulated. However, such an explanation seems inconsistent with findings from studies reporting biphasic N400/P600 patterns for combined manipulations (e.g. Kutas & Hillyard, 1980; Kolk et al., 2003; Hagoort, 2003; Hoeks, Stowe, & Doedens, 2004; also see van Petten & Luka, 2012). Alternatively, one could seek an explanation for the apparent inconsistencies in the signal itself. That is, rather than producing a qualitatively different modulation pattern, plausibility and semantic association could combine quantitatively: while the manipulation of plausibility increases P600 amplitude, the manipulation of semantic association results in a negativity, thus attenuating the P600.

More specifically, the unit of measurement underlying the ERP signal is a single scalar representing the scalp-recorded voltage at a given electrode and point in time (for a given subject and experimental trial), and any recorded voltage could in principle reflect the combined contribution of multiple relevant factors. Standard approaches to analyzing ERPs (e.g. using ANOVAs or linear mixed models), however, model signal variance at the voltage level, and hence do not allow for assessing such a quantitative combination of factors. Indeed, in order to investigate if and how different factors combine in the signal, one needs to isolate their independent contributions to the recorded voltages. This is, however, non-straightforward (Brouwer & Crocker, 2017).

Here, we show how rERP estimation (Smith & Kutas, 2015a, 2015b) extends standard approaches to ERP analysis by adding a layer of signal variance modeling that does allow for assessing how relevant, manipulated factors combine quantitatively in the signal. The core idea is to replace every single observed, scalp-recorded voltage by a regression-based estimate, which derives from a linear combination of factors relevant to the experimental manipulations. More specifically, the idea is to first fit linear regression models to the voltage-level data, in order to split the observed scalar voltages into the contributions made by these factors at the underlying level; that is, at the level at which they combine into the scalp-recorded voltage. Henceforth, we will refer to this underlying level as the *latent voltage* level. The resultant models then allow for the regression-based estimation of rERP waveforms, while offering strict control over the individual factors that contribute to it. Hence, they offer a means of assessing how factors quantitatively combine at the latent voltage level. Crucially, the resultant *regression-based* rERP waveforms are no different in nature than the ERP waveforms based on observed voltages, and can therefore be plotted and analyzed using standard analysis tools (e.g., ANOVAs or linear mixed models).

In what follows, we will first introduce the rERP framework, and discuss how it extends traditional approaches towards ERP analysis with an additional layer of signal variance modeling. Next, we will incrementally derive an rERP analysis of the DBC results, and show how this reveals spatiotemporal component overlap between the N400 and the P600, thereby offering an explanation for the WCS-derived inconsistencies in terms of its underlying LCS. Finally, we discuss the implications of our results for past and future studies as well as for neurocognitive theories.

2 | REGRESSION-BASED ERP ESTIMATION

In order to model the how experimental manipulations combine at the latent voltage level, we will employ rERP estimation (Smith & Kutas, 2015a, 2015b). The core idea is to replace each individual voltage measurement in the ERP data—each observed voltage scalar—with a voltage estimate from a linear regression model that optimally combines the manipulated variables to explain the variance in the signal. The resultant rERP data can then be analyzed in a similar manner as the original ERP data, but crucially, it can also be analyzed when manipulations are controlled for (kept constant), thus allowing one to examine how the manipulations combine at the latent voltage level, and how this in turn affects the voltage-level averaged signal, that is, the rERP waveforms.

Smith and Kutas (2015a) motivate the ERP framework from the observation that the traditional method of ERP estimation through averaging is a specific instance of general least squares regression, and hence that the full power of linear (and non-linear) regression can be harnessed to estimate ERPs. That is, an ERP is the stimulus-locked neural activity caused by post-synaptic potentials. When recorded for a single trial (within a subject) this activity will include stimulus-evoked potentials, as well as background activity that is not related to the stimulus. ERP estimation through averaging, then, is grounded in the idea that across trials $y_1 \dots y_n$, the stimulus-locked ERP β is stable (i.e. the brain's systematic response to the stimulus of interest), whereas the background noise ε_i is random. That is, if one averages over trials, $\varepsilon_1 \dots \varepsilon_n$ will cancel each other out, and the ERP β will be isolated. For a given electrode (e.g. Pz) and latency (e.g., 400 ms), the ERP β is thus estimated as the average of the scalp-recorded potentials $y_1 \dots y_n$:

$$\frac{1}{n}\sum_{i}^{n} y_i = \beta, \tag{1}$$

where n equals the number of trials within a given experimental condition and subject.

Crucially, Smith and Kutas (2015a) show that this averaging procedure can be recast as linear regression, such that the scalp-recorded potential at each electrode, latency, and trial y_i is estimated as the ERP β plus by-trial varying noise ε_i:

$$y_i = \beta + \varepsilon_i, \tag{2}$$

while y_i is known, both β and ε_i are not, and hence neither can be solved for. Least squares regression, however, allows for the estimation of β by minimizing the total squared noise $\left(\sum_i^n \varepsilon_i^2\right)$. Crucially, it can then be shown that the best estimate of β is in fact the mean of $y_1 \ldots y_n$ as estimated in Equation (1) (see Smith & Kutas, 2015a, for a derivation). Indeed, this means that ERP averaging is just a special instance of least squares regression, implying that one can replace the simplistic regression model in Equation (2) with a more general least squares regression model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i. \quad (3)$$

Here, y_i is again the voltage measured for given a trial, electrode and latency, and ε_i is again the by-trial varying noise. The ERP part of the right-hand side, however, is now broken down into a sum of weighted predictors. Each predictor x_{ji} represents a specific property of the stimulus presented at trial i,[2] typically reflecting experimental manipulations. Predictors may be either categorical (e.g. stimulus felicity) or continuous (e.g. probability).

By finding the set of β_j's that minimizes total squared noise across trials, the ERP at this electrode and latency thus effectively gets decomposed into a linear combination of the predictors x_{ji}.[3] In a fitted regression model, these predictors can then be set to values representing the properties of a specific trial stimulus, such that computing $\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$ gives the predicted voltage for that trial at a given electrode and latency. When computed for each subject, electrode, and latency, one obtains an rERP data set, in which each observed voltage is replaced by an estimated voltage that derives from a linear combination of weighted predictors encoding stimulus properties. Crucially, to the extent that the resultant rERP data closely matches the ERP data—which can be assessed by examining voltage residuals and variance (see below)—an analysis of how the predictors combine to produce the voltage estimates provides a window into how experimental manipulations combine at the latent voltage level.

It is important to note that the regression models are not explicitly exposed to any *categorical* condition-coding predictors, but rather to predictors encoding stimulus-related properties of the individual trials. That is, within each subject, a regression model is fitted over trials for each electrode and latency, and a voltage estimate is then computed for each trial from this fitted model. From these estimated voltages, rERPs can then be computed in the same way ERPs are computed from observed voltages. To keep the distinction between rERPs and ERPs clear, we will henceforth follow Smith and Kutas (2015a) in using the label rERPs to refer to waveforms derived from averaging estimated voltages, and ERPs to refer to waveforms derived from averaging observed voltages. Although regression models are not exposed to condition-coding predictors, trials can be regrouped by condition after fitting and estimation is completed, such that rERPs, like ERPs, can be compared across conditions, and subjected to standard statistical analyses. Figure 1 gives a schematic comparison between the ERP and rERPs approach.[4]

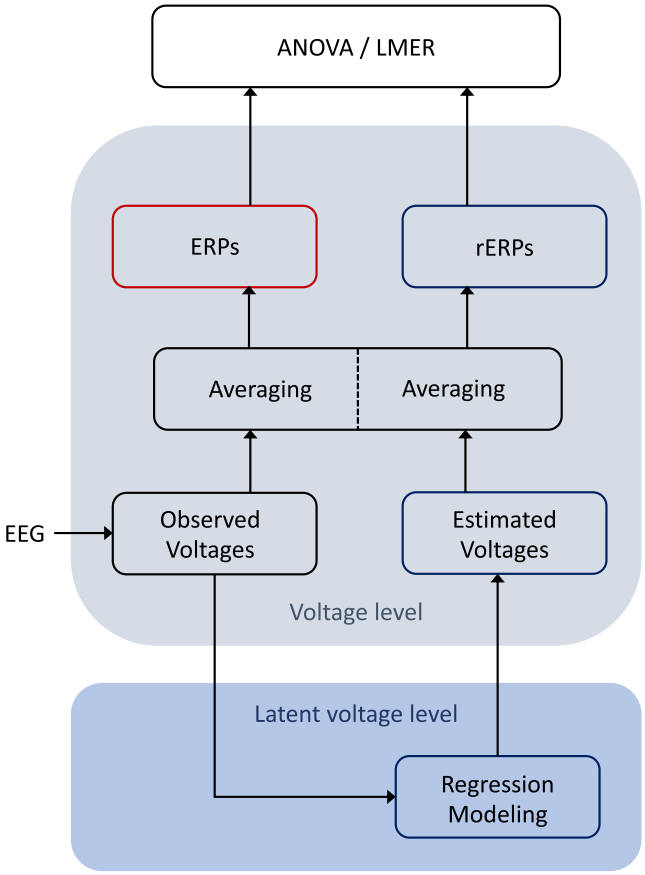

FIGURE 1 Schematic comparison of event-related potentials (ERPs) and rERPs. Both approaches start from the scalp-recorded voltages. The ERP approach averages these observed voltages directly, and subjects the resultant averages to statistical analysis. Hence, the ERP approach only involves voltage-level modeling of the signal. The rERP approach, by contrast, first employs regression modeling to replace each observed voltage by a voltage estimate that derives from a linear combination of relevant, experimentally manipulated factors. It is these voltage estimates that are then averaged, and subjected to statistical analysis. The rERP approach thus adds a layer of latent voltage-level signal modeling

[2]This in fact is no different from Equation (2) which is shorthand for $y_i = \beta_1 x_{1i} + \varepsilon_i$ with $x_{1i} = 1$ for all trials.

[3]Although β_j's are no longer found through simple averaging, they can still be efficiently estimated using standard techniques.

[4]Please refer to the electronic version for color figures.

In sum, the rERP framework extends traditional approaches to ERP analysis by offering a means to isolate the independent contributions of relevant factors at the latent voltage level, and to examine how they quantitatively combine into the voltage-level rERP signal. In what follows, we will demonstrate how this additional layer of statistical modeling allows one to offer a quantitative explanation for the puzzling results from the Delogu et al. (2019) experiment, by reconciling apparent inconsistencies in voltage-level modulations of the ERP waveforms at the underlying, latent voltage level.

3 | REGRESSION-BASED ESTIMATION OF DBC

Delogu et al. (2019, DBC) conducted an ERP experiment on language comprehension, in which they manipulated the *plausibility* and semantic association (henceforth *association*) of a target word in German mini-discourses, across three conditions:

Baseline [+plausible, +associated]
Johann betrat das Restaurant. Wenig später öffnete er die Speisekarte und […]
'*John entered the restaurant. Before long, he opened the menu and* […]'
Event-related [−plausible, +associated]
Johann verließ das Restaurant. Wenig später öffnete er die Speisekarte und […]
'*John left the restaurant. Before long, he opened the menu and* […]'
Event-unrelated [−plausible, −associated]
Johann betrat die Wohnung. Wenig später öffnete er die Speisekarte und […]
'*John entered the apartment. Before long, he opened the menu and* […]'

In this context manipulation design, the critical word in the sentence (e.g., "Speisekarte"/"menu") rendered the entire mini-discourse either plausible (baseline) or implausible (event-related and event-unrelated). Figure 2 (right) shows the plausibility ratings (1–7 point scale) obtained from 30 participants. Moreover, the critical word was either associated with a prime word in the first sentence (e.g. "restaurant" in the baseline and event-related conditions) or not ("Wohnung"/"apartment" in the event-unrelated condition). Figure 2 (right) shows the prime-target association ratings (1–7 point scale) obtained from 20 participants.

The ERP results are shown in Figure 3. DBC found that the event-related condition—in which only plausibility was low—produced a P600 effect relative to baseline in the 600–1,000 ms time-window. Indeed, consistent with the literature this seems to suggest that the plausibility manipulation affects the processes underlying the P600 component of the ERP signal (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007, for reviews). However, in the event-unrelated condition—in which both plausibility and association were low—an N400 effect was observed in the 300–500 ms time-window, which continued as sustained

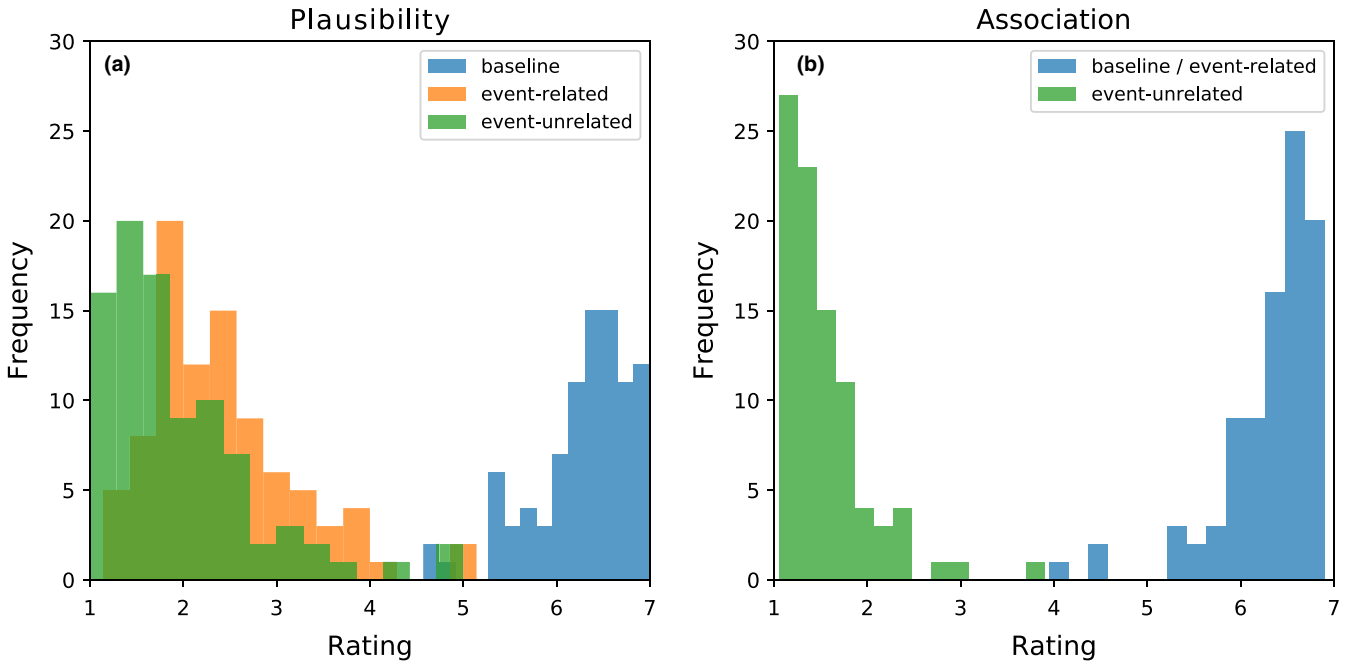

FIGURE 2 Distributions of the Plausibility (left) and Association (right) ratings by condition for the stimuli of the DBC experiment

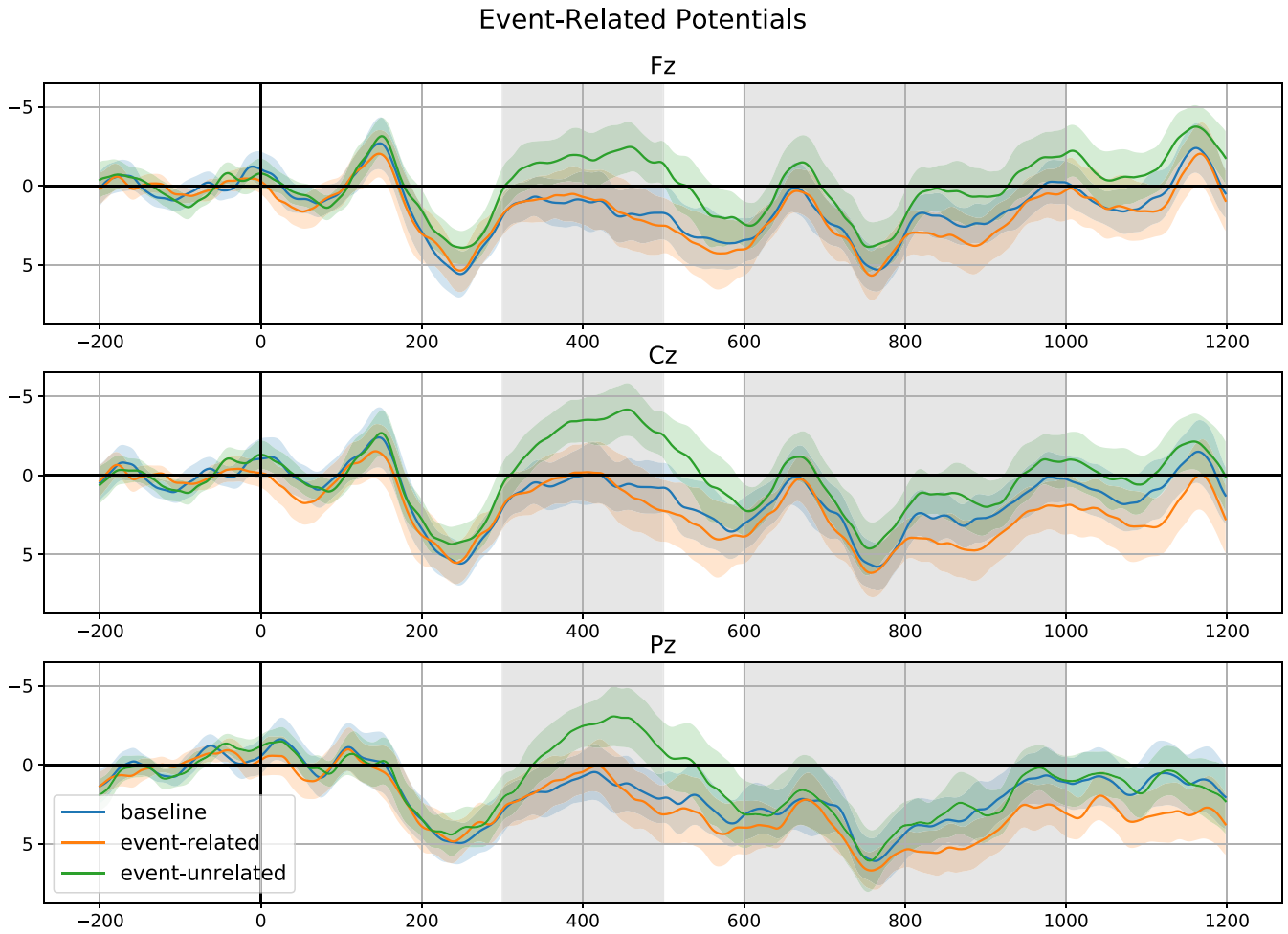

FIGURE 3 Grand-average event-related potential waveforms for the DBC experiment. Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

negativity until the end of the ERP epoch (–1,200 ms), but crucially, no clear P600 effect, although a small but significant positivity was present at occipital sites. The apparent absence of a P600 effect appears inconsistent with the pattern of results for the event-related condition, and requires an explanation (see van Petten & Luka, 2012, for more instances of such conflicting results).

One possibility is that the manipulation of plausibility affects processing—and thereby the ERP signal—differently depending on whether it co-occurs with a manipulation of association (event-unrelated) or not (event-related). This implies an architectural explanation, and while possible, none of the extant neurocognitive theories of the N400 and the P600 in language processing predicts such a pattern of findings (see Delogu et al., 2019, for discussion). Another possibility, by contrast, is that the plausibility and association manipulations do both independently affect processing, but that these manipulations combine at the latent voltage level, thereby rendering it unclear what is going on in the ERP waveforms derived from voltage-level averaging. More specifically, it is possible that plausibility and association have an opposite influence on the ERPs: While decreasing plausibility drives waveforms to go more positive in the 600–1,000 ms P600 time-window, leading to an increase in P600 amplitude for the implausible but associated (event-related) trials relative to baseline, lower association simultaneously exerts an even stronger pull in the negative direction, yielding a net sustained negativity for implausible and unassociated (event-unrelated) trials. Below, we will explore this latter explanation of the data by incrementally deriving an rERP analysis of the DBC results.

3.1 | Establishing a baseline—Modeling no difference between conditions

A first step towards an rERP analysis is to establish a baseline account of the variance in the signal. Assume that no factors were systematically manipulated, and that all trials effectively belonged to the same condition. The optimal voltage estimate for each subject, electrode and time point, would

then simply be the average over trials $y_1 \ldots y_n$, which in regression terms is equivalent to an intercept-only model (see Section 3):

$$y_i = \beta_0 + \varepsilon_i. \quad (4)$$

Here, the intercept β_0 is the expected mean of y across trials (and conditions). Regression models are fitted for each electrode ($N = 18$: Fz, Cz, Pz, F3, FC1, FC5, F4, FC2, FC6, P3, CP1, CP5, P4, CP2, CP6, O1, Oz, and O2), and within each subject ($N = 21$) and time point ($N = 700$, given a sample rate of 500 Hz and ERP epochs lasting from −200 to 1,200 ms), yielding a total of 26,400 models. If one estimates voltages from a regression model for a given electrode, time point, and subject, the same voltage estimate will thus result for each trial (and hence each condition). Figure 4 shows the rERP waveforms resulting from these voltage estimates.

As the intercept-only model predicts the expected mean of y for each trial, it predicts no differences between trials. The intercept-only model does thus not adequately account for the differences in variance between conditions (all conditions in Figure 4 fully overlap), and hence offers a poor fit to the data. The differences between the observed and estimated voltages, the residuals, quantify this fit: The closer this difference is to 0, the better the fit of the estimates to the observed voltages. These residuals, shown in Figure 5, reveal that in the 300–500 ms N400 time-window, there is error for trials from each condition: event-unrelated (implausible and unassociated) trials should be more negative, while event-related (implausible but associated) and baseline trials should be more positive. Moreover, in the 600–1,000 ms P600 time-window, the intercept-only model fails to explain the more sustained negative going deflection for the event-unrelated trials, as well as the more positive deflection for event-related trials. This does, however, provide us with a clear baseline to improve upon; that is, the aim is to explain away the error of the intercept-only model by extending it with predictors that encode stimulus properties that were experimentally manipulated in the DBC experiment: plausibility and association.

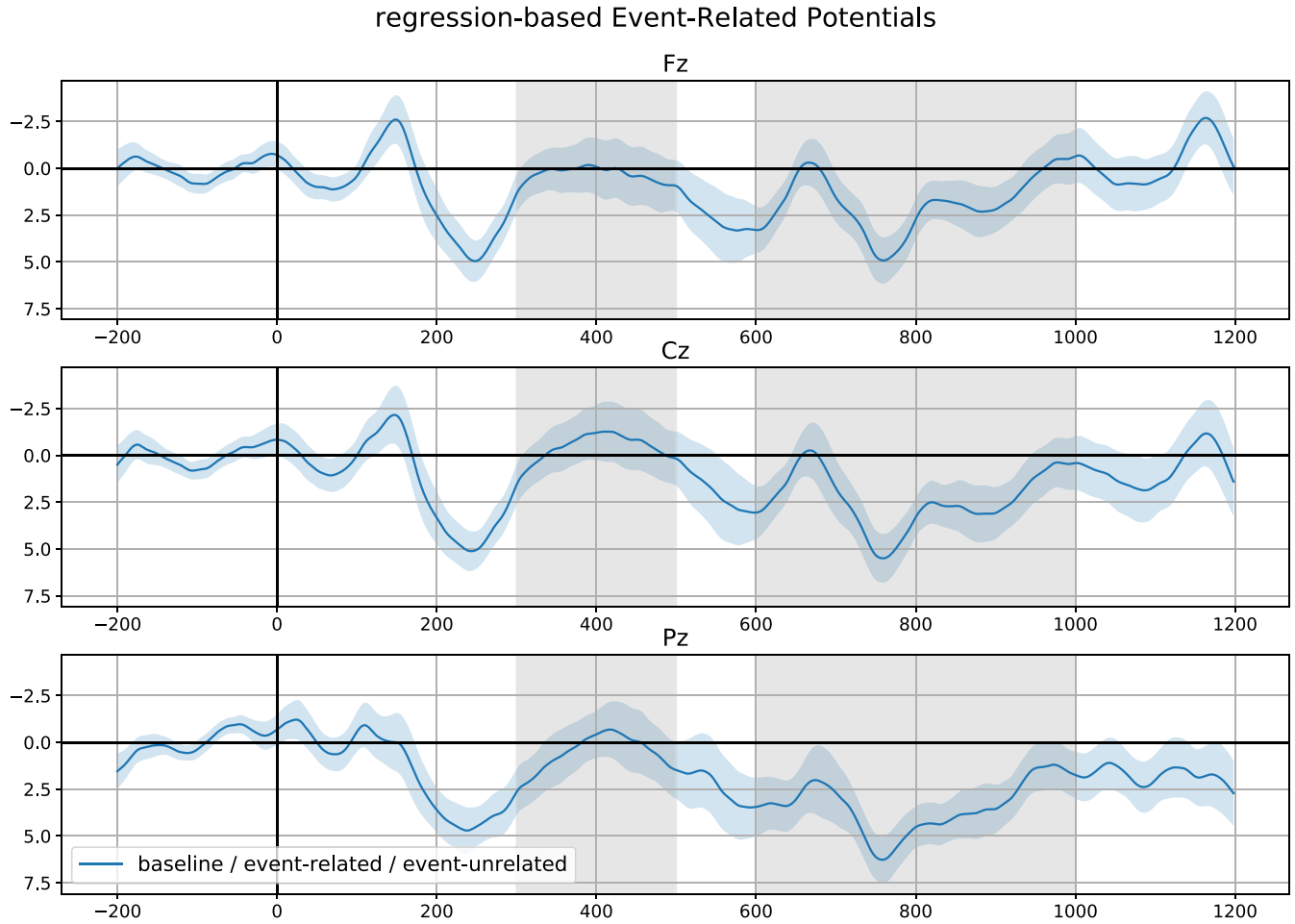

FIGURE 4 Grand-average regression-based event-related potential waveforms resulting from the intercept-only model (estimated voltage y from Equation 4). Note that only one line is shown as the predictions for the different conditions fully overlap. Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

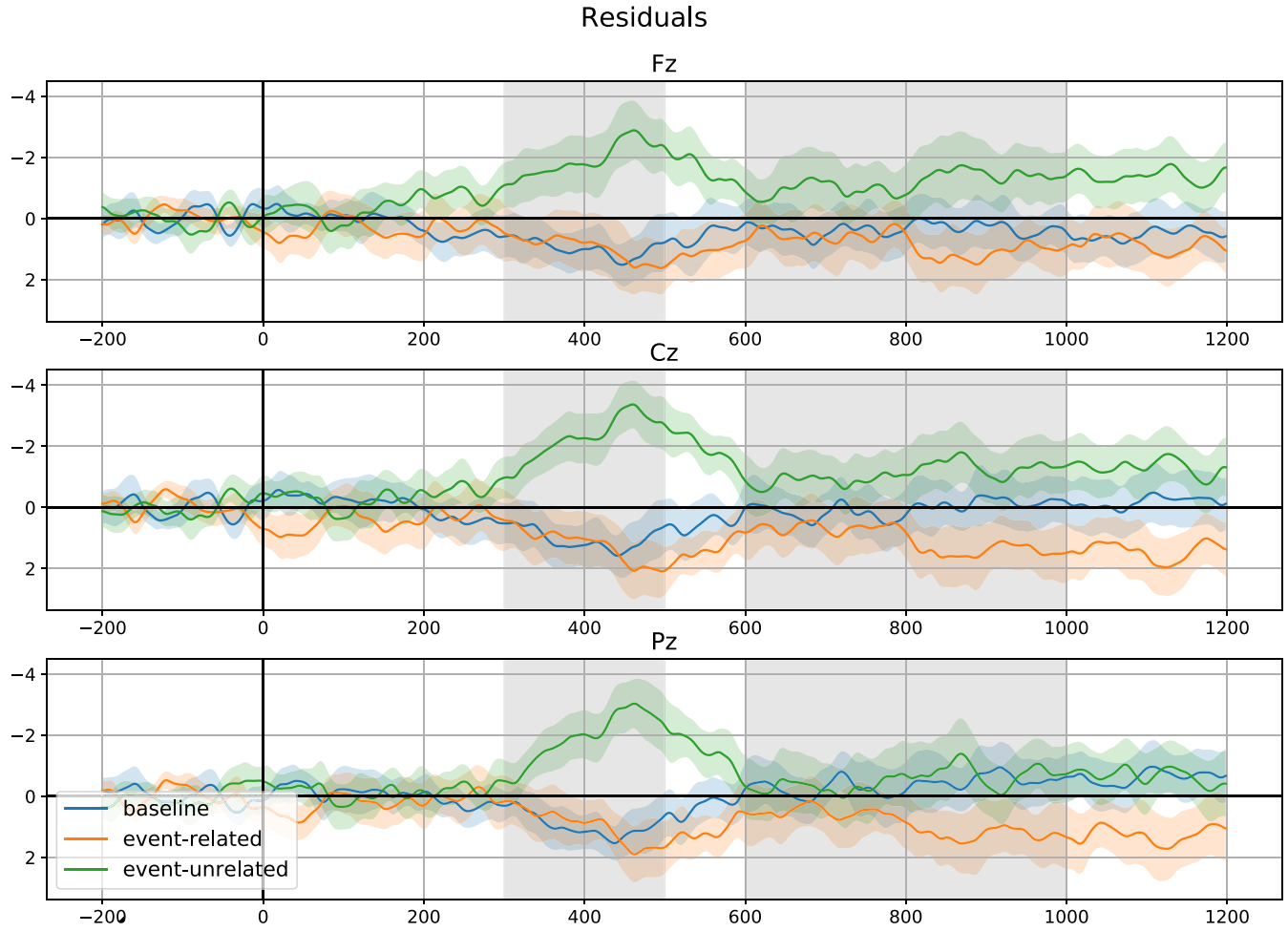

FIGURE 5 Grand-average residuals between the observed voltages and the voltages estimated from the intercept-only model. Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

3.2 | Modeling the influence of plausibility

DBC manipulated plausibility such that trials from the event-related and event-unrelated condition were more implausible than trials from the baseline condition (see Figure 2). Plausibility ratings were collected offline for each item, meaning that by-trial plausibility can be entered as a predictor into an rERP analysis. Prior to entering this predictor into the models, however, two transformations are applied to the plausibility ratings:

1. First, in order to render the interpretation of the slopes more intuitive in relation to the ERP components, the scale of the plausibility ratings is inverted by subtracting each rating (which was expressed on a 1–7 point scale) from the maximum possible rating (7). As a result, higher ratings now indicate more implausible trials, whereas lower ratings indicate more plausible trials;
2. Subsequently, these inverted ratings are z-transformed, such that a rating of 0 indicates mean (im)plausibility, while negative ratings reflect plausible, and positive ratings implausible trials.

Entering the transformed plausibility ratings as a predictor into the intercept-only model (4), yields the following regression model:

$$y_i = \beta_0 + \beta_1 \text{ plausibility } + \varepsilon_i. \quad (5)$$

Figure 6 shows the resulting rERP waveforms. Two things immediately stand out. First, in the 300–500 ms N400 time-window, both implausible conditions (event-related and event-unrelated) are more negative than the plausible one (baseline). Second, there are no large differences in the 600–1,000 ms P600 time-window, other than a slight frontal–posterior negative–positive gradient for the implausible conditions relative to baseline.

While it is unsurprising that the rERPs deriving from the plausibility-only model do not match the ERPs observed by DBC, it is informative to see how the voltage estimates underlying these rERPs arise. Figure 7 shows the residuals for the plausibility-only model. These residuals reveal that the model neatly captures the variance for the baseline trials across the entire epoch. They further reveal, however, that from the onset of the N400 time-window onwards, it fails to

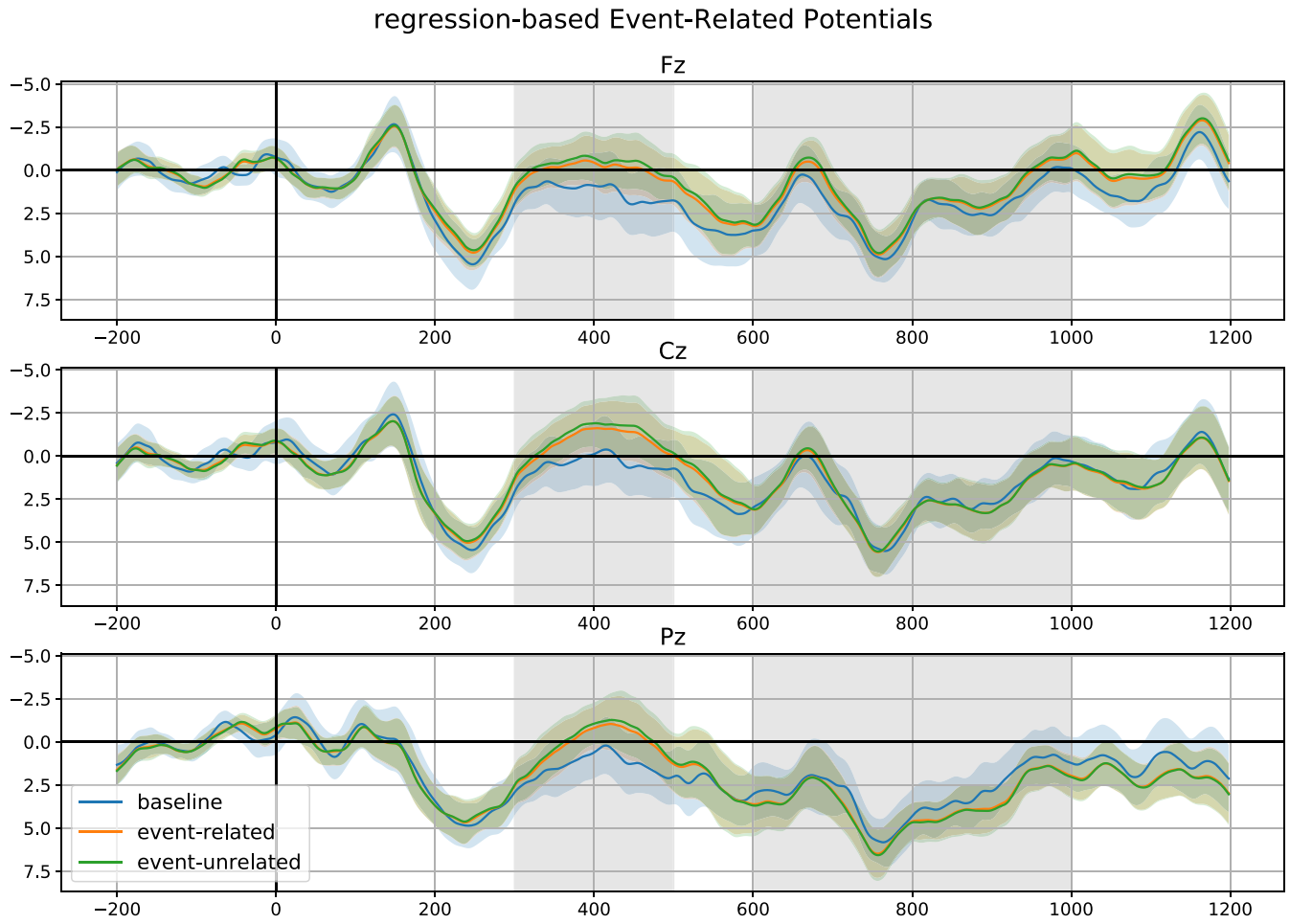

FIGURE 6 Grand-average regression-based event-related potentials resulting from the intercept plus plausibility model (estimated voltage y from Equation 5). Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

account for the more negative going voltages in the event-unrelated trials, as well as for the more positive going voltages in the event-related trials. Let us unpack how this leads to the rERPs depicted in Figure 6. Starting with the increased negativity for the implausible conditions, half of the implausible trials that the model is exposed to occur with a larger negative deflection (event-unrelated trials) relative to the other half of the implausible trials (event-related trials), which match the baseline trials in the N400 time-window. Hence, the optimal voltage estimates for implausible trials (minimizing $\sum_i^n \varepsilon_i^2$) are ones that are intermediate to the voltages observed for event-related and event-unrelated trials (compare mean amplitudes within the highlighted N400 time-window in Figures 3 and 6). Note that the slight difference between the implausible conditions (event-unrelated > event-related) can be attributed to the fact that event-unrelated conditions were rated as slightly more implausible than event-related conditions (see Figure 2, left). Indeed, this is exactly why the residuals show that the event-unrelated trials should be more negative, while the event-related trials should be more positive. A similar explanation applies to the absence of any differences, other than the frontal–posterior negative–positive gradient for implausible conditions, in the 600–1,000 ms P600 time-window. Again, half of the implausible trials (event-unrelated) are more negative than plausible trials (baseline), while the other half of the implausible trials (event-related) are more positive. Hence, optimal voltage estimates for implausible trials are again estimates that are intermediate to the observed voltages, which in this case neatly align with the voltages observed for the plausible trials. This is clearly reflected in the residuals as well, which again show that the event-unrelated trials should be more negative, while event-related trials should be more positive. In sum, plausibility alone does thus not suffice to explain the DBC results.

3.3 | Modeling the influence of association

DBC also manipulated association, such that the critical word in the baseline and event-related trials were semantically associated with the prior context, while this was not

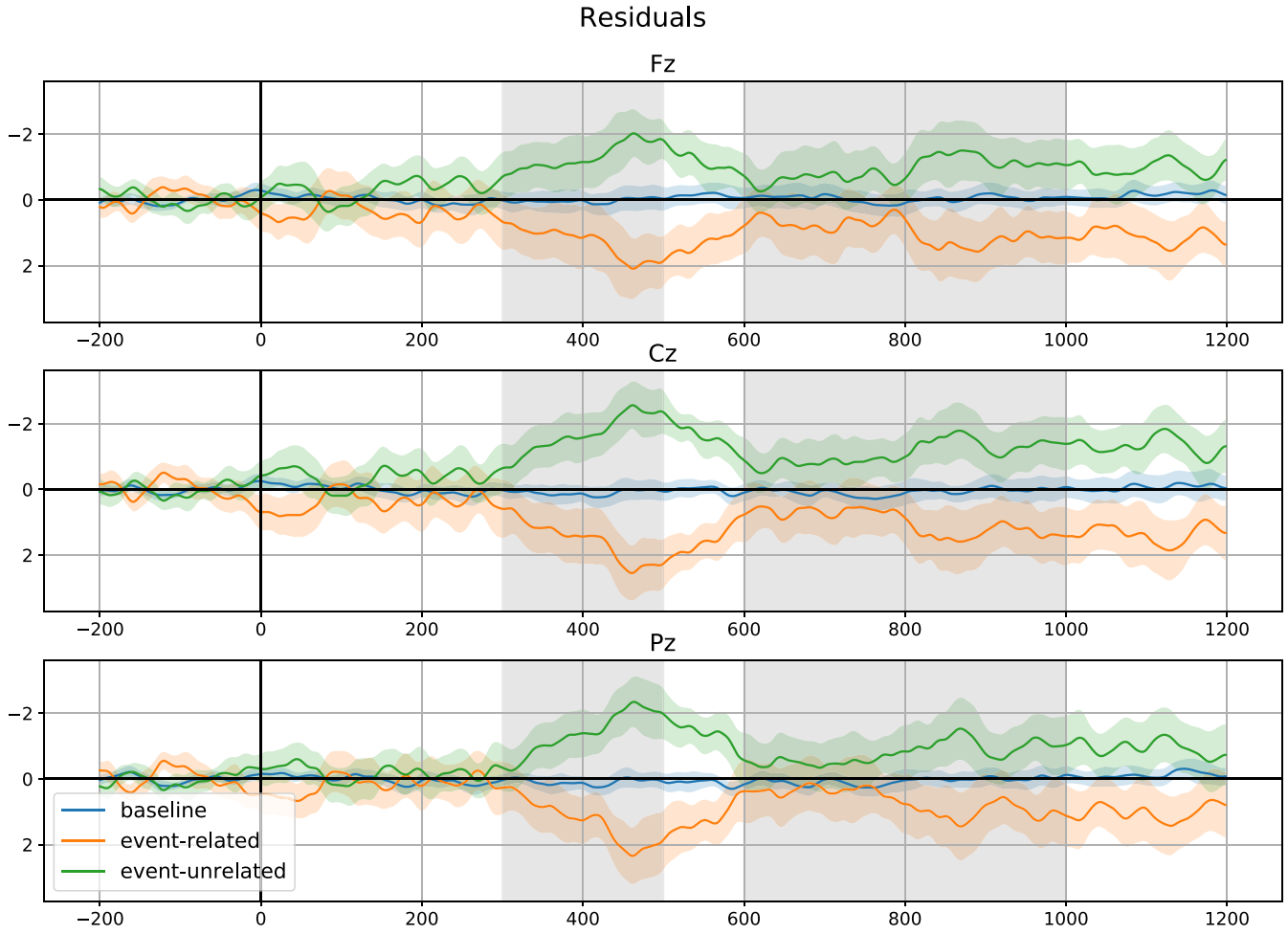

FIGURE 7 Grand-average residuals between the observed voltages and the voltages estimated from the intercept plus plausibility model. Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

the case in the event-unrelated trials (see Figure 2). As with plausiblity, association ratings were collected offline for each item, and hence by-trial association ratings can be entered as a predictor into an rERP analysis. Prior to doing so, the same transformations that were applied to the plausibility ratings are also applied to the association ratings:

1. The scale of the association ratings is inverted by subtracting each rating (which was expressed on a 1–7 point scale) from the maximum possible rating (7). As a result, lower ratings now indicate more associated trials, whereas higher ratings indicate less associated trials;
2. Subsequently, these inverted ratings are z-transformed, such that a rating of 0 indicates mean association, while negative ratings reflect more associated, and positive ratings less associated trials.

To examine the influence of association independent of plausibility, we again first extend the intercept-only model (4) with association as a predictor, yielding the following model:

$$y_i = \beta_0 + \beta_1 \text{ association } + \varepsilon_i. \quad (6)$$

Figure 8 shows the rERP waveforms resulting from this model. While association alone is enough to adequately explain the sustained negative deflection for the event-unrelated relative to the baseline trials, it fails to explain increased positivity for event-related relative to baseline trials. Indeed, this pattern was to be expected, given that only event-unrelated trials differ in association. The residuals, shown in Figure 9, show that while the estimated voltages of the model are very close to the observed voltages for the event-unrelated trials with low association, there is considerable error for event-related and baseline trials, which have equally high association, especially in the 600–1,000 ms P600 time-window. That is, estimates for the baseline trials should be more negative, while estimates for the event-related baseline should be more positive. Indeed, the estimated voltages for these trials are again intermediate to the observed voltages, as the model cannot distinguish between these conditions on the basis of association alone.

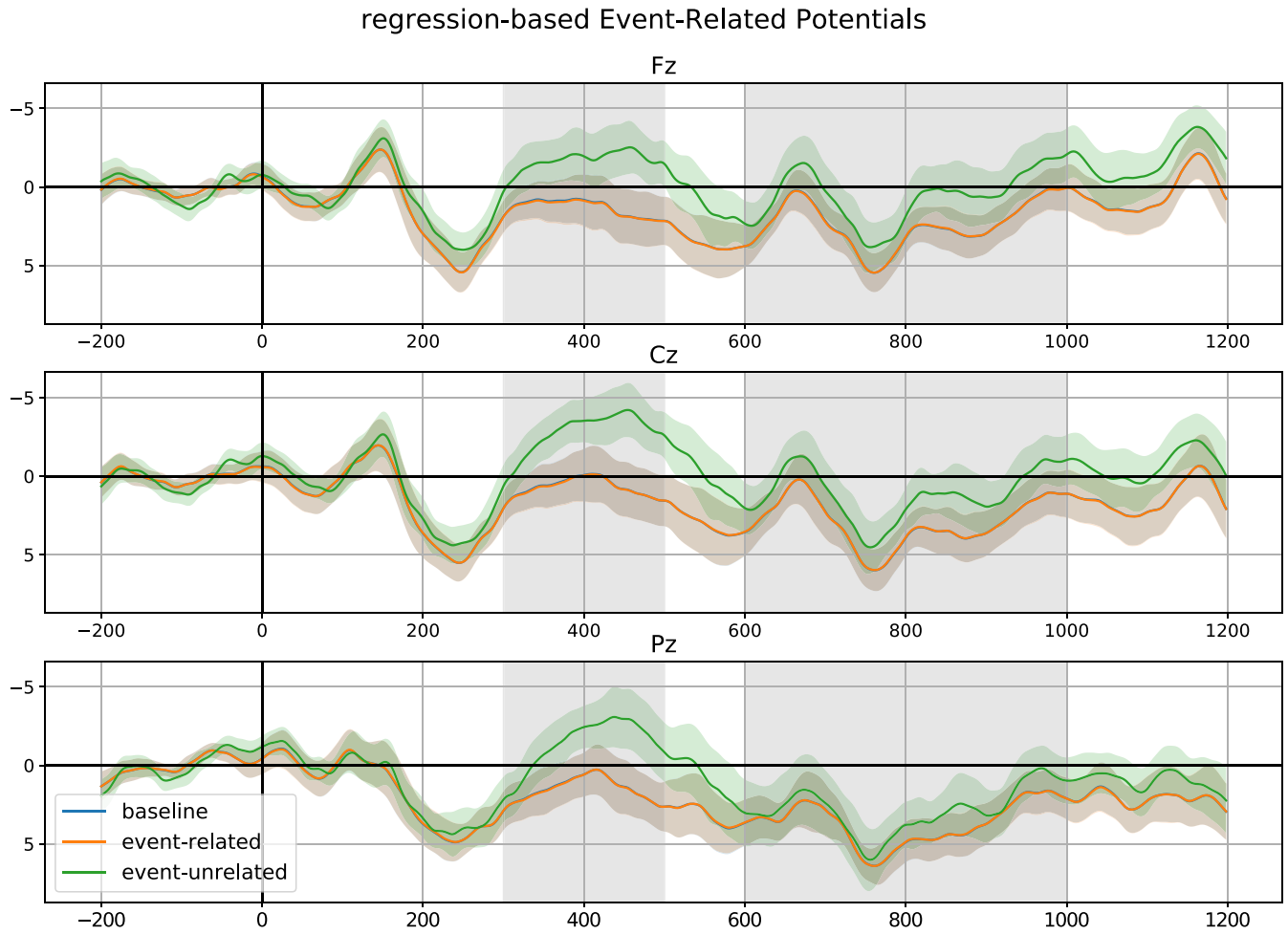

FIGURE 8 Grand-average regression-based event-related potentials resulting from the intercept plus association model (estimated voltage y from Equation 6). Note that the baseline condition is not visible, as the predictions for the event-unrelated and baseline trials overlap. Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

3.4 | Modeling the combined influence of plausibility and association

The analyses above establish that neither plausibility, nor association alone are sufficient to adequately account for the observed variance in the signal across trials. Crucially, however, trials differed on combinations of both of these dimensions, raising the question if an rERP analysis that includes both plausibility and association as predictors could adequately explain the data.

Entering both predictors into the rERP analysis, yields the following model:

$$y_i = \beta_0 + \beta_1 \text{ plausibility } + \beta_2 \text{ association } + \varepsilon_i. \quad (7)$$

Figure 10 shows the rERP waveforms resulting from this model. Visual inspection clearly shows that this model provides a better fit to the data than the previous analyses. First, in the 300–500 ms N400 time-window, it captures the negativity for the event-unrelated trials (relative to the baseline and the event-related trials). Second, in the 600–1,000 ms P600 time-window it captures both the sustained negativity for the event-unrelated (relative to baseline), as well as the positivity for the event-related trials (relative to baseline).

While this qualitative fit looks promising, the next step is to quantify goodness of fit. Like in the previous analyses, we can turn to the residual voltages to see how close the voltage estimates of the models are to the observed voltages. Figure 11 shows the residuals for the current model. Given that over the entire epoch, the mean residuals for each condition are much closer to 0 than in the previous analyses, it can be concluded that the current rERP analysis offers the best quantitative fit to the data. An open question, however, remains whether the rERP data leads to the same effect structure as the ERP data. To investigate this, we can subject the rERP data to the same statistical analyses as the ERP data.

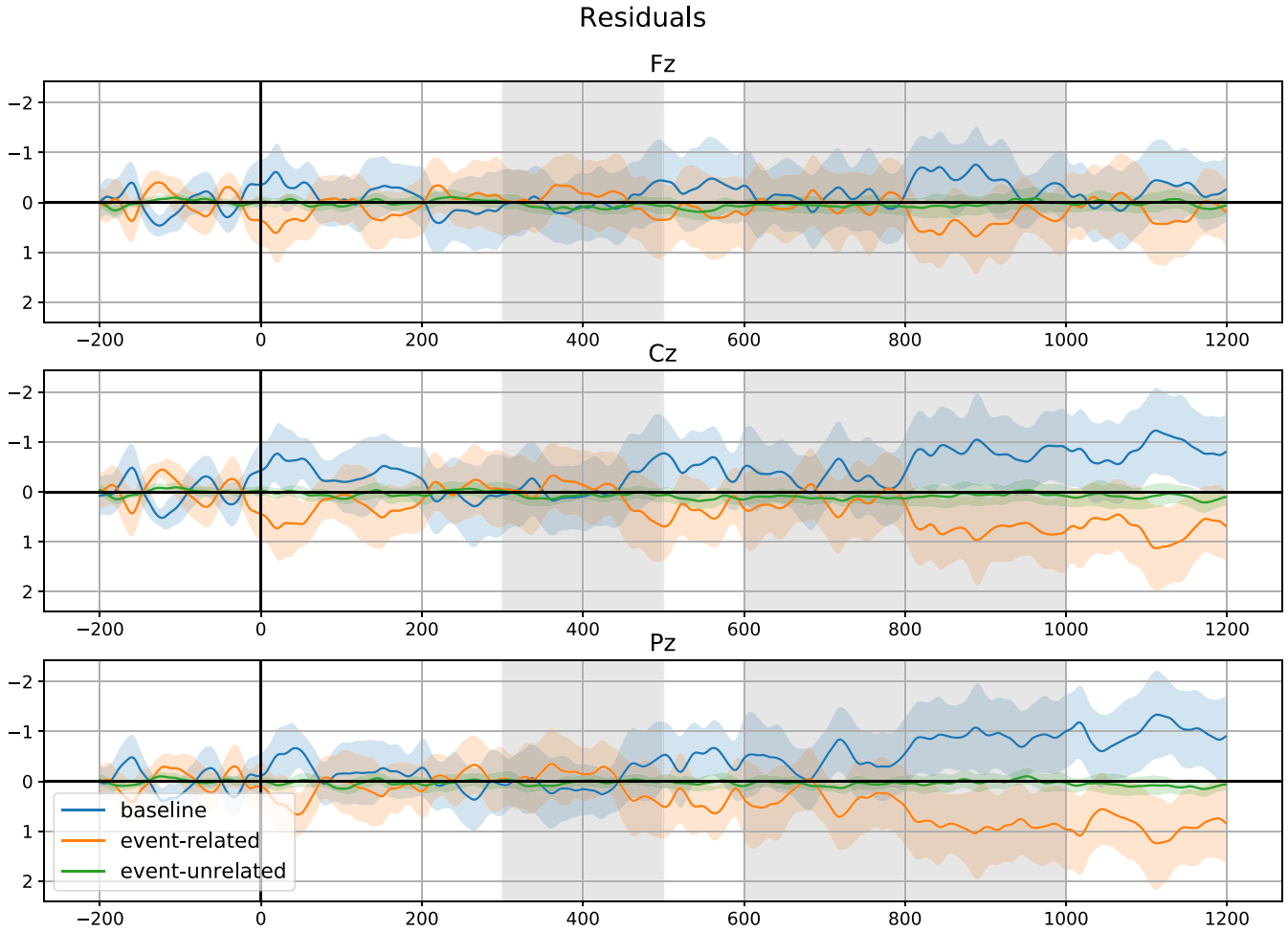

FIGURE 9 Grand-average residuals between the observed voltages and the voltages estimated from the intercept plus association model. Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

3.4.1 | Statistical analyses

Following DBC, we computed mean amplitudes for each condition and electrode in the 300–500 ms (N400) and 600–1,000 ms (P600) time-window from the estimated voltages. To examine the topographic distribution of the effects, data from midline and lateral electrodes were treated separately. Data from midline sites included three electrodes (Fz, Cz, Pz). Data from lateral sites were grouped into four regions of interest (ROIs): left anterior (F3, FC1, FC5), right anterior (F4, FC2, FC6), left posterior (P3, CP1, CP5), and right posterior (P4, CP2, CP6). Within each time-window, ANOVAs over midline electrodes were carried out with Condition (baseline, event-related, event-unrelated) and anterior–posterior (AP) distribution (anterior, central, posterior) as repeated measure factors. The ANOVAs over later sites included Condition, AP distribution (anterior, posterior) and Hemisphere (left, right) as within-subject factors. The Greenhouse–Geisser correction was applied to all ANOVAs with greater than one degree of freedom in the numerator. In such cases, the corrected *p*-value is reported. Generalized *eta*-squared ($\eta^2 G$) is reported as a measure of effect size.

N400 time-window (300–500 ms)

The ANOVA on midline electrodes revealed a significant effect of Condition, $F(2, 40) = 23.97$, $p < .001$, $\eta^2 G = 0.19$. No other effects or interactions were significant (all $Fs < 12$). As shown in Table 1, the difference between event-unrelated ($M = -1.95$; $SD = 2.7$) and baseline ($M = 0.98$; $SD = 2.96$) was significant, but the difference between event-related ($M = 1.36$; $SD = 3.17$) and baseline was not.

The ANOVA on lateral sites revealed an effect of Condition, $F(2, 40) = 22.42$, $p < .001$, $\eta^2 G = 0.198$. No other effects or interactions were significant (all $Fs < 1$). As shown in Table 1, the difference between event-unrelated ($M = -1.64$; $SD = 2.27$) and baseline ($M = 0.87$; $SD = 2.43$) was again significant, while the difference between event-related ($M = -1.06$; $SD = 2.73$) and baseline was not.

Finally, to further assess the fit of the rERPs to the ERPs, we re-ran the ANOVAs with Type (rERP, ERP) as a

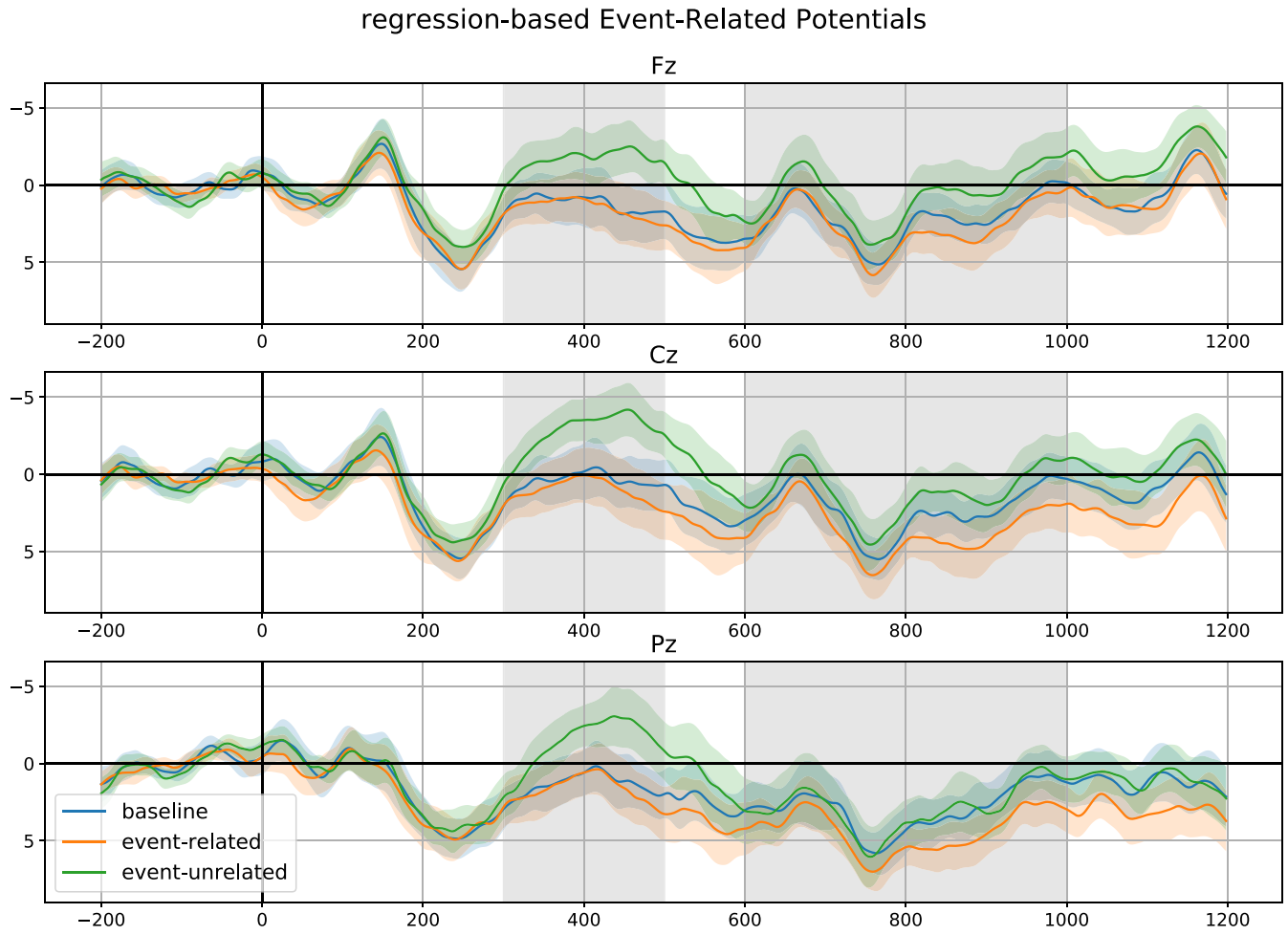

FIGURE 10 Grand-average regression-based event-related potentials resulting from the intercept plus plausibility plus association model (estimated voltage *y* from Equation 7). Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

between-subjects factor. Both the ANOVAs on midline electrodes, as well as the ANOVAs on lateral sites, revealed no significant effect of Type (all $Fs < 1$).

P600 time-window (600–1,000 ms)

The ANOVA on midline sites revealed a significant effect of Condition, $F(2, 40) = 6.59$, $p < .005$, $\eta^2G = 0.07$, and an interaction of Condition and AP distribution, $F(4, 80) = 3.02$, $p < .04$, $\eta^2G = 0.009$. As shown in Table 1, the comparison between event-unrelated ($M = -1.59$; $SD = 2.99$) and baseline ($M = 2.51$; $SD = 2.49$) showed a significant interaction of Condition and AP distribution, indicating a more pronounced negativity for the event-unrelated condition over more anterior sites (see Figure 10), while the comparison between event-related ($M = 3.55$; $SD = 2.52$) and baseline revealed a significant effect of Condition.

The ANOVA on lateral sites showed an effect of Condition, $F(2, 40) = 4.82$, $p < .02$, $\eta^2G = 0.05$, and an interaction of Condition and AP distribution, $F(2, 40) = 5.14$, $p < .02$, $\eta^2G = 0.008$. As shown in Table 1, the comparison of both event-unrelated ($M = 1.51$; $SD = 2.65$) and even-related ($M = 3.03$; $SD = 2.46$) with baseline ($M = 2.29$; $SD = 2.17$) revealed a significant interaction of Condition and AP distribution, indicating a more anterior negativity for event-unrelated and a more posterior positivity for event-related.

Again, to further assess the fit of the rERPs to the ERPs, we re-ran the ANOVAs with Type (rERP, ERP) as a between-subjects factor. Both the ANOVAs on midline electrodes, as well as the ANOVAs on lateral sites, revealed no significant effect of Type (all $Fs < 1$).

In summary, the rERP data neatly replicate the ERP data reported by DBC. In the N400 time-window, implausible and unassociated (event-unrelated) trials elicited an N400 effect compared to the baseline condition, while no such N400 effect was observed for implausible but associated (event-related) trials relative to baseline. In the P600 time-window, in turn, the rERPs revealed a sustained negativity for implausible and unassociated (event-unrelated) trials relative to baseline, while a P600 effect was observed for implausible but associated (event-related) trials compared

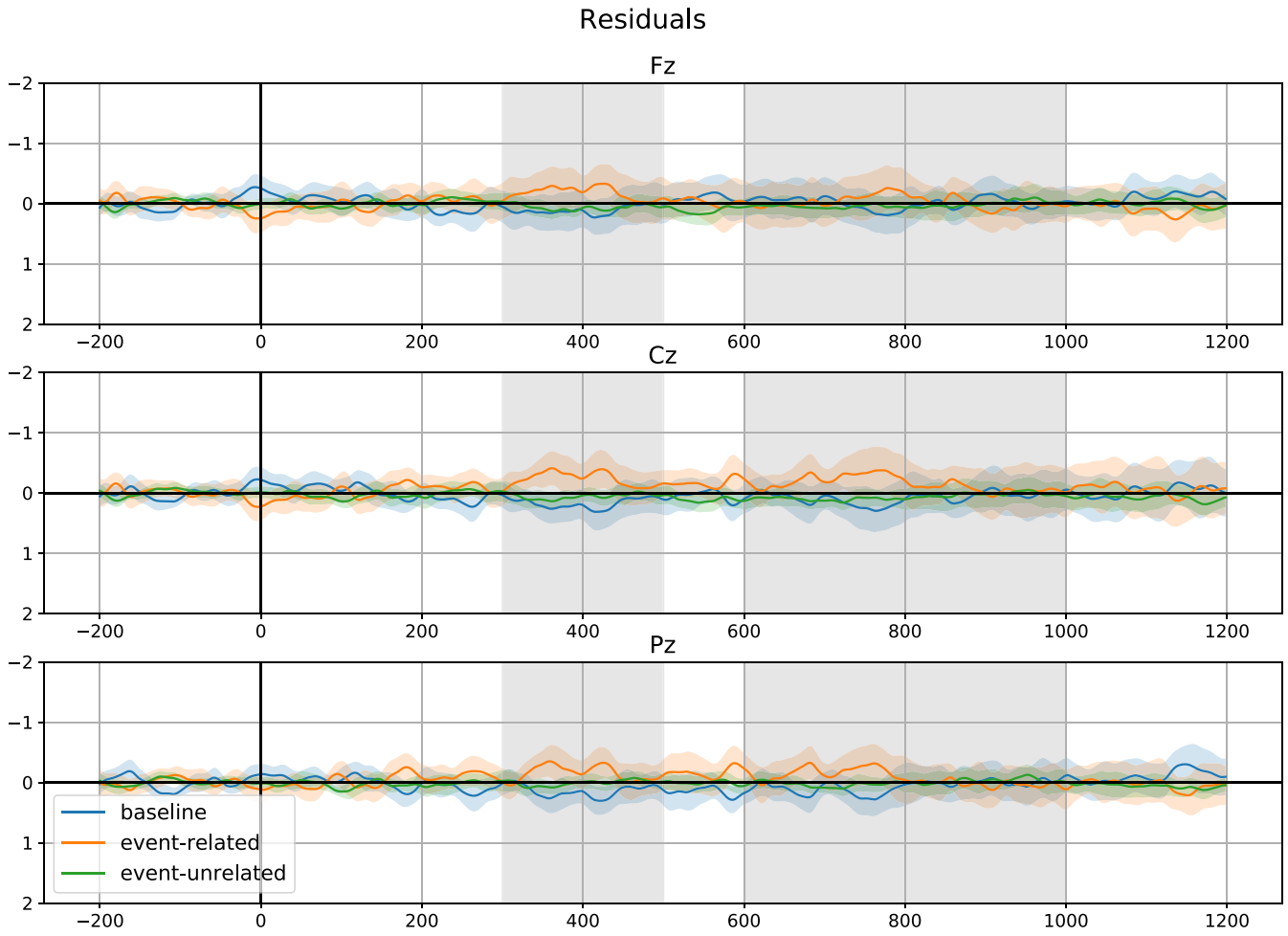

FIGURE 11 Grand-average residuals between the observed voltages and the voltages estimated from the intercept plus plausibility plus association model. Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

TABLE 1 ANOVAs on rERPs to target nouns across the N400 time-window and the P600 time-window

			300–500 ms			600–1,000 ms		
		df	*F*	*p*	η^2G	*F*	*p*	η^2G
Event-related vs baseline								
Midline	Cond	(1, 20)	<1	.47	0.003	5.56	.03	0.04
	Cond × AP	(2, 40)	<1	.56	<0.001	2.05	.16	0.005
Lateral	Cond	(1, 20)	<1	.68	0.001	2.83	.11	0.02
	Cond × AP	(1, 20)	<1	.71	<0.001	5.16	.03	0.006
	Cond × H	(1, 20)	<1	.60	<0.001	<1	.89	<0.001
Event-unrelated vs baseline								
Midline	Cond	(1, 20)	40.1	<.001	0.19	2.65	.12	0.02
	Cond × AP	(2, 40)	<1	.41	0.001	5.15	.02	0.011
Lateral	Cond	(1, 20)	41.6	<.001	0.20	2.70	.12	0.02
	Cond × AP	(1, 20)	<1	.88	<0.001	10.1	<.01	0.01
	Cond × H	(1, 20)	<1	.88	<0.001	<1	.89	<0.001

Notes: Cond × AP = Condition × Anterior–Posterior distribution; Cond × H = Condition × Hemisphere.

to baseline. Crucially, none of the ANOVAs with Type as between-subjects factor, revealed a significant difference between the rERPs and ERPs. Hence, beyond offering a close quantitative fit, as evidenced by the voltage-level residuals (see Figure 11), the rERP data do thus also adequately mimic ERP data in terms of variance.

3.4.2 | How plausibility and association combine

Up to this point, we have added the rERP layer, and established that the resultant voltage estimates provide a close fit to the observed voltages, both in terms of residuals, as well as in terms of variance. The aim of adding this layer of analysis was to obtain a means to split the observed scalar voltages into the individual contributions made by plausibility and association. The fitted regression coefficients offer precisely that: A means to examine how these factors individually contribute to the scalp-recorded voltages.

Figure 12 plots the fitted coefficients over time. Note that while the intercept (β_0) is shown as is, the coefficients for the predictors are "anchored" to the intercept to aid interpretability (plausibility: $\beta_0 + \beta_1$; association: $\beta_0 + \beta_2$). The coefficients reveal two things. First, in the 300–500 ms N400 time-window, the increased negativity for the unassociated (event-unrelated) trials is entirely driven by association; that is, the coefficients for association are negative (and hence have a negative offset from the intercept in Figure 12), such that they yield more negative voltages when trials are less associated, while the coefficients for plausibility are near-zero (and hence close to the intercept in 12), meaning that plausibility does not affect the voltage estimates in this time-window. The N400 effect observed for event-unrelated relative to baseline (and event-related) trials thus seems to be driven by the difference in association. Second, in the 600–1,000 ms P600 time-window, the coefficients for association and plausibility pull in opposite direction: the coefficients for association remain negative (and keep their negative offset to from the intercept in Figure 12), again yielding more negative voltage estimates for less associated trials, while the coefficients for plausibility are positive (as reflected by a positive offset from the intercept in Figure 12), yielding more positive voltage estimates for more implausible trials. This straightforwardly explains the P600 effect for the event-related condition relative to baseline: Implausible, but associated trials

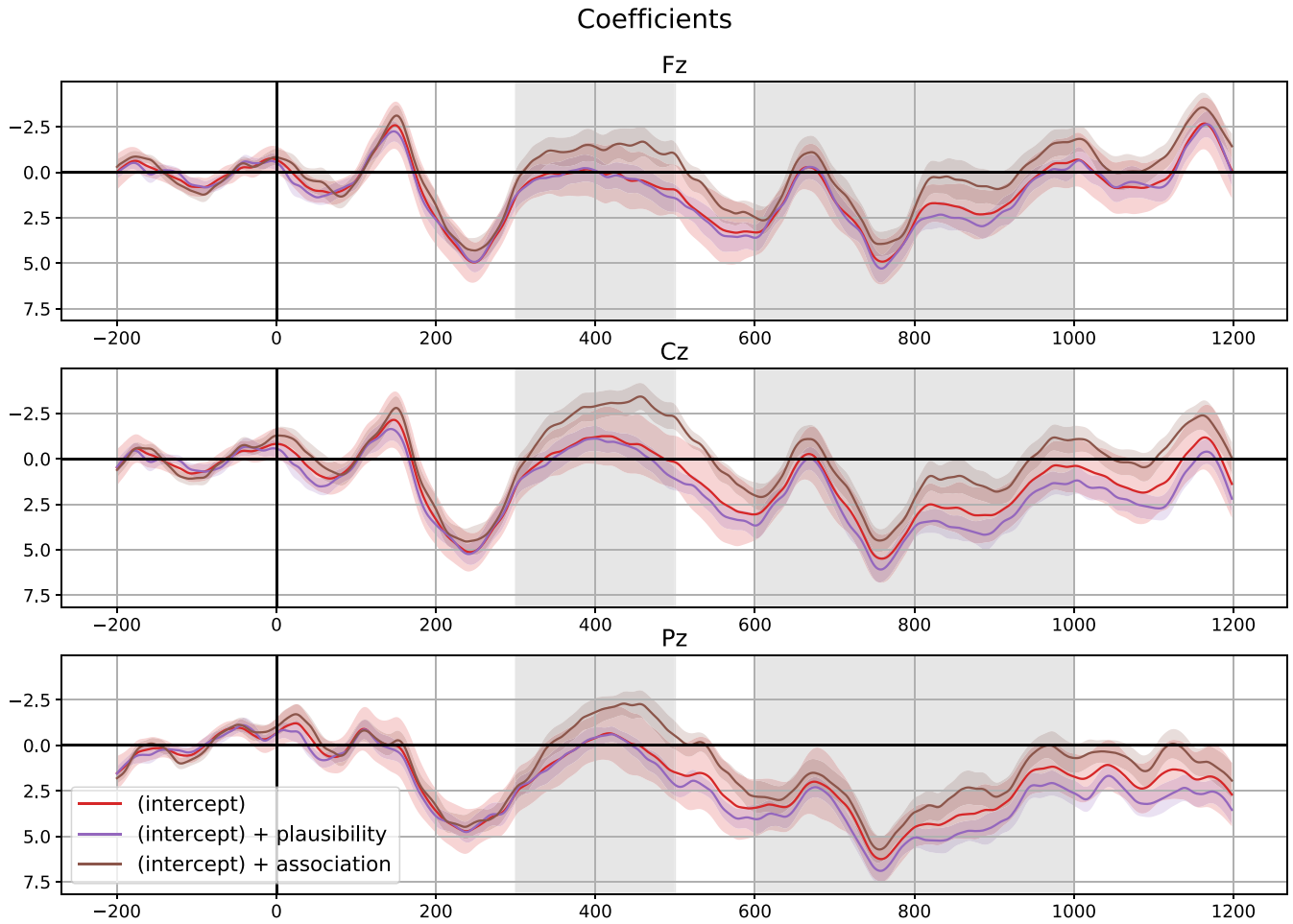

FIGURE 12 Grand-average coefficients from the intercept plus plausibility plus association model. Slopes of predictors are "anchored" to the intercept (e.g. β_1 is plotted as $\beta_0 + \beta_1$; see text for details). Negative is plotted upwards. Shaded regions show mean voltage ±2 *SE* across subjects

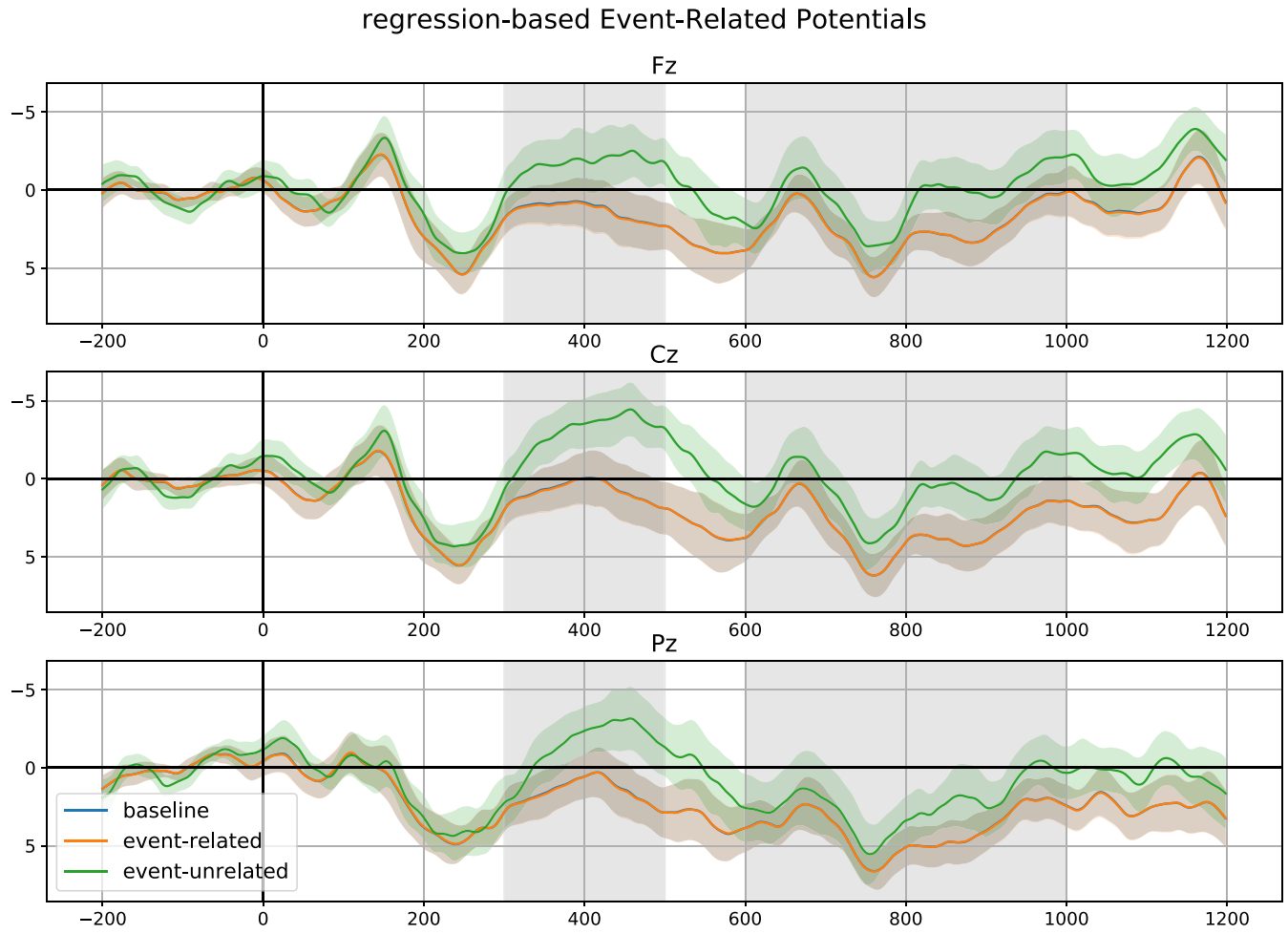

FIGURE 13 Grand-average regression-based event-related potentials resulting from the intercept plus plausibility plus association model when plausibility is set to its mean rating (0) for all trials (estimated voltage y from Equation 8). Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

produce an increased positivity relative to baseline trials, which is driven by the difference in plausibility. The explanation for the sustained negativity for the event-unrelated condition relative to baseline, by contrast, is more complex, as it arises from the independent quantitative contribution of both plausibility and association. That is, while plausibility yields more positive voltage estimates for the implausible-unassociated trials, association yields more negative voltage estimates. Crucially, as the absolute coefficients for association are larger than the absolute coefficients for plausibility, association is the stronger force: For the event-unrelated trials, which have both low plausibility and low association, the sustained negativity thus arises as the negative pull of association overrules the positive pull of plausibility, thereby producing a negativity relative to baseline. Indeed, as event-related trials have low plausibility but high association, this negative pulling force is not exerted for these trials, hence producing an increased positivity relative to baseline, driven purely by lower plausibility.

The influence of association independent of plausibility

The rERP analysis allows us to examine how plausibility and association combine in more depth by keeping the influence of one (or more) predictor(s) constant. One can, for instance, isolate the influence of association by setting plausibility to its mean (0; as the predictor is z-transformed) for all trials in the fitted models:

$$y_i = \beta_0 + \beta_1 0 + \beta_2 \text{ association } + \varepsilon_i. \quad (8)$$

Note that we do not need to re-fit the models, rather we use the fitted models to re-estimate the rERP waveforms. Figure 13 shows the rERP waveforms from model (7), with mean plausibility for all trials (8). First, this confirms that the N400 effect in the 300–500 ms time-window is driven by association. Second, it also shows the that the sustained negativity in the 600–1,000 ms time-window does indeed arise from a quantitative combination between plausibility and association; that is, by neutralizing the influence of plausibility, the sustained negativity now gets overestimated (compare Figures 3 and 13).

The influence of plausibility independent of association

We can also keep association constant by setting it to its mean (0) for all trials in the fitted models:

$$y_i = \beta_0 + \beta_1 \text{ plausibility } + \beta_2 0 + \varepsilon_i. \quad (9)$$

Figure 14 shows the re-estimated rERP waveforms from model (7), with mean association for all trials (9). By neutralizing the influence of association, two things become apparent. First, we lose the increased negativity in the 300–500 ms N400 time-window for event-unrelated trials relative to baseline (and event-related) trials, as only the coefficients of association have an effect in this time-window, and association is now the same for all trials. Second, both event-related and event-unrelated trials now yield an increased positivity in the 600–1,000 ms P600 time-window (note that for the event-related trials, this positivity is overestimated due to neutralizing the negative pull of association; compare Figures 3 and 14). Indeed, this suggests that plausibility is reflected in an increase in positivity. Moreover, these positivities already start to emerge at 400 ms, that is, in the middle of the N400 time-window. We will return to this latter observation in the discussion.

Statistical analysis (see above) confirms this pattern:

N400 time-window (300–500 ms)

ANOVAs on midline and lateral sites revealed no significant effects or interactions (all $Fs < 1$).

P600 time-window (600–1,000 ms)

ANOVAs on midline electrodes revealed a significant effect of Condition, $F(2, 40) = 5.33$, $p < .03$, $\eta^2 G = 0.04$. No further effects or interactions were significant. As shown in Table 2, both event-unrelated ($M = 2.95$; $SD = 2.41$) and event-related ($M = 2.87$; $SD = 2.31$), were significantly different from baseline ($M = 1.83$; $SD = 2.73$).

ANOVAs on lateral sites revealed a significant interaction between Condition and AP distribution, $F(2, 40) = 5.13$, $p < .04$, $\eta^2 G = 0.006$. As shown in Table 2, the comparison of both event-unrelated ($M = 2.55$; $SD = 2.30$) and event-related ($M = 2.51$; $SD = 2.19$) with baseline ($M = 1.78$; $SD = 2.38$)

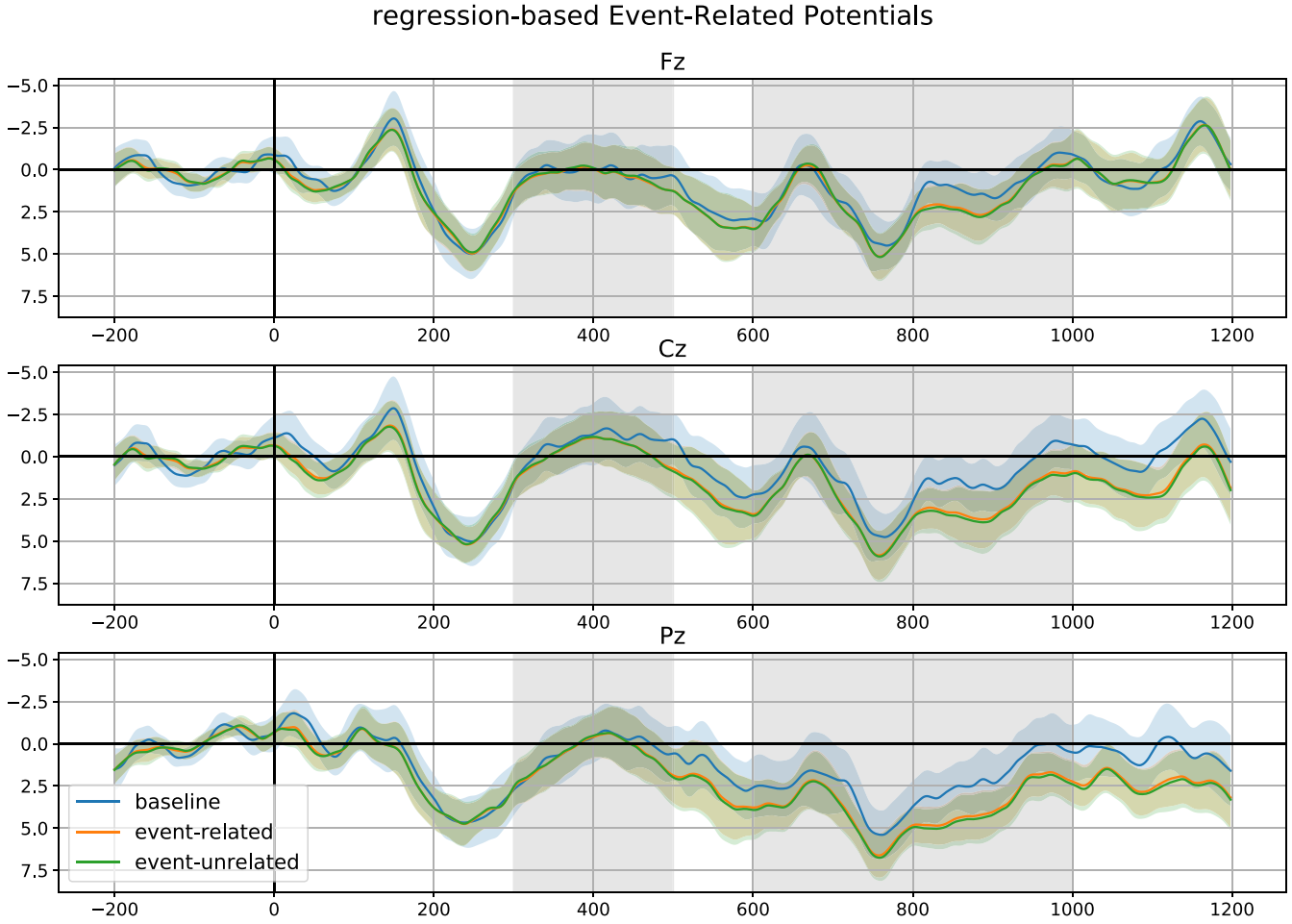

FIGURE 14 Grand-average regression-based event-related potentials resulting from the intercept plus plausibility plus association model when association is set to its mean rating (0) for all trials (estimated voltage y from Equation 9). Negative is plotted upwards. Shaded regions show mean estimated voltage ±2 *SE* across subjects

TABLE 2 ANOVAs on rERPs to target nouns across the P600 time-window: the influence of Plausibility

			600-1,000 ms		
		df	*F*	*p*	η^2G
Event-related vs baseline					
Midline	Cond	(1, 20)	5.65	.03	0.035
	Cond × AP	(2, 40)	2.08	.15	0.005
Lateral	Cond	(1, 20)	2.85	.11	0.02
	Cond × AP	(1, 20)	4.94	.04	0.006
	Cond × H	(1, 20)	<1	.91	<0.001
Event-unrelated vs baseline					
Midline	Cond	(1, 20)	5.16	.03	0.04
	Cond × AP	(2, 40)	2.18	.14	0.006
Lateral	Cond	(1, 20)	2.57	.12	0.02
	Cond × AP	(1, 20)	5.27	.03	0.007
	Cond × H	(1, 20)	<1	.93	<0.001

Notes: Cond × AP = Condition × Anterior–Posterior distribution; Cond × H = Condition × Hemisphere.

produced a significant interaction with AP distribution, indicating a more pronounced effect over posterior sites.

In summary, statistical analysis confirms absence of an N400 effect and presence of a P600 effect for both the event-unrelated and the event-related conditions compared to baseline.

3.5 | Discussion

We have presented an rERP analysis of the DBC data that extends the standard voltage-level approach to ERP analysis by adding a layer of analysis that allows for splitting the observed scalar voltages into the contributions made by the relevant, experimentally manipulated factors: plausibility and association. Our rERP analysis was validated by showing that the voltage estimates closely match the observed voltages, both in terms of residuals, as well as in terms of variance. Crucially, the fitted coefficients of the regression models reveal that plausibility and association combine quantitatively in producing the voltage estimates: While a lower degree of association tends to pull the estimated voltages to be more negative, from the 300 ms until the end of epoch, a lower degree of plausibility pulls the estimated voltages to be more positive, from about 400 ms until the end of the epoch. Indeed, the fact that association and plausibility pull in opposite direction from about 400 ms onwards, offers a quantitative, signal-based explanation for why DBC did not observe a positivity for implausible and unassociated (event-unrelated) trials. Even though the implausibility of these trials leads the coefficients for plausibility to produce a positive shift in the estimated signal, the coefficients for association simultaneously produce a negative shift, and as the absolute coefficients for association are larger than the absolute coefficients for plausibility, this leads to the net effect of a sustained negativity. Moreover, it was shown that if the influence of association is controlled for, by setting association with its mean across all trials, we do observe an increased positivity in the rERPs for event-unrelated trials in the 600–1,000 ms P600 time-window, while we no longer observe any negativity in the 300–500 ms N400 time-window. Overall, the rERP analysis thus clearly supports the conclusion that 1) association is reflected in N400 amplitude (and that in the DBC data this negativity sustains into the 600–1,000 ms time-window), 2) plausibility is reflected in P600 amplitude, and 3) association (N400) and plausibility (P600) combine quantitatively in producing the rERP waveforms (see Figures 10 and 12), most notably in the 600–1,000 ms P600 time-window.

How does this rERP analyis map onto the neural generator dynamics underlying the ERP signal? Indeed, the rERP analysis suggests that the N400 and the P600 overlap in time, with the plausibility-driven P600 emerging in the 300–500 ms N400 time-window (see Figure 14), and the association-driven N400 sustaining into the 600-1,000 ms P600 time-window (see Figure 13). When mapped onto neural generator dynamics, the implication would be that the processes, and hence the generators, underlying the N400 component and the P600 component are active simultaneously, which in turn leads to spatiotemporal overlap between the N400 and the P600 in the scalp-recorded ERP signal. Brouwer and Crocker (2017) have recently pointed out that if such spatiotemporal component overlap is at play, the standard Waveform-based Component Structure (WCS) approach to ERPs—which derives component structure (e.g., the modulation pattern of the N400 and the P600) by looking at effects on mean amplitude in predetermined time-windows—may lead to the spurious presence or absence of effects, and consequently to inconsistent data patterns. In fact, they argue that taking mean amplitudes in these time-windows as indicative of components is incorrect, as at any given point a waveform may merely show the summation of the *latent* components contributing to the ERP signal at that time (see also Donchin et al., 1978; Duncan et al., 2009; Luck, 2005a, 2005b; Näätänen, 1982; Squires et al., 1975). Crucially, given that components index specific computational processes (Näätänen & Picton, 1987), and that these processes may temporally overlap, multiple components may be contributing to the scalp-recorded ERP signal at any point in time. The observed WCS is thus nothing more than an epiphenomenon of its underlying LCS, and due to spatiotemporal component overlap, WCS and LCS may look very different.

Brouwer and Crocker (2017) argue that apparent WCS-derived inconsistencies within (e.g. Delogu et al., 2019; Kim & Osterhout, 2005) and across (see van Petten & Luka, 2012) studies, may be resolved at the LCS level. They also point out,

however, that investigating LCS is non-trivial, as scalp-recorded voltage scalars inherently conflate the contributions of multiple latent components. Crucially, we have here shown how the rERP framework—as proposed by Smith and Kutas (2015a, 2015b)—extends the standard voltage-level analysis approaches to ERPs, by providing the mathematical tools to split these voltage scalars into the independent contributions made by different, relevant experimental factors, thereby offering a powerful means to investigate LCS.

4 | GENERAL DISCUSSION

Traditional analyses of ERPs examine how experimental conditions modulate the pattern of observed ERP components as reflected by the mean amplitude in predetermined time-windows. Crucially, this Waveform-based Component Structure (WCS) approach often leads to inconsistent results within (e.g. Delogu et al., 2019; Kim & Osterhout, 2005; Kolk et al., 2003; Kuperberg et al., 2007) as well as across studies (see Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg, 2007; van Petten & Luka, 2012, for reviews). Motivated by the observation that such WCS-derived inconsistencies may be reconciled by factoring in spatiotemporal overlap between ERP components in the LCS underlying the WCS (Brouwer & Crocker, 2017), we have here shown how regression-based ERP (rERP; Smith & Kutas, 2015a, 2015b) estimation allows for the explicit modeling of LCS using linear regression. Crucially, analysis of the resultant regression models allows one to derive an explanation for the WCS in terms of how relevant regression predictors combine in space and time, and crucially, how individual predictors may be mapped onto unique components, revealing how these spatiotemporally overlap in the WCS.

The rERP approach effectively extends traditional approaches towards ERP analysis by adding an additional layer of analysis that replaces each observed scalp-recorded voltage with a regression-based estimate, which decomposes this voltage into the contribution made by different experimentally manipulated factors. For the rERP analysis to be valid, it is essential that the differences between the observed voltages and the estimated voltages, the residuals, are as close to zero as possible. That is, the observed scalp-recorded voltages represent the true signal and the rERP analysis should adequately capture this signal. Indeed, the closer the estimated voltages are to the observed voltages, the closer the rERP waveforms will be to the ERP waveforms. In fact, for any rERP analysis that closely fits the observed data, there is no fundamental difference between ERPs and rERPs; that is, it does not matter if one averages the observed or the estimated voltages to obtain a waveform, or if one carries out statistical analyses on the ERPs or rERPs. Indeed, the only difference between ERPs and rERPs is that rERPs derive from one additional step of statistical estimation of the signal (prediction and averaging), as compared to ERPs (averaging only). Crucially, the main advantage of the rERP layer is that valid rERP analyses with multiple relevant predictors allow one to investigate how these predictors combine by keeping one (or more) predictor(s) constant. Indeed, while the resultant voltage estimates may then look very different from the observed voltages, these estimates still derive from the true signal as recorded from the scalp. As such, rERPs derived from averaging these estimated voltages are again just as valid a statistical estimate of the signal as the ERPs derived from averaging the observed voltages, and hence they can also be treated as and subjected to the same statistical analysis as normal ERPs.

The requirement for a valid rERP analysis to minimize residuals highlights the fact that we effectively harness the additional layer of regression-based modeling as a "machine learning" layer rather than an "inferential statistics" layer. That is, we employ linear regression modeling to estimate voltages from a set of experimentally relevant predictors, and these estimates are then processed in the same way as observed voltages are treated in traditional approaches to ERP analysis. Indeed, in the present paper we have only looked at how individual predictors affect the overall rERP waveforms, and we have conducted inferential statistics at the level of these waveforms by looking for effects on mean amplitude in predetermined time-windows. In theory, one could also turn to regression models for inferential statistics by looking at the relative importance of predictors within a given model. Indeed, while this could be done at the level of the regression models fitted for each subject and time point, one would ideally construct a single model for each point in time and conflate over subjects. To this end, linear mixed effects regression (LMER) could be employed to account for between-subject (and optionally between-item) variability (see Nieuwland et al., 2020, for such an approach).

While this approach is valid, and could potentially be insightful, the core requirement of the rERP approach still holds: the difference between the observed and estimated voltages should be minimal. That is, if the fit between the regression-based estimates and the scalp-recorded voltages is not assessed, one is at risk of drawing unwarranted conclusions about the importance of predictors: a predictor could be highly significant in a model that fits the observed voltages very poorly, potentially leading to wrong conclusions about that predictor. For instance, Delogu et al. (2019) also collected by-item Cloze ratings for their target words (measured as the fraction of sentence completions using a specific word), and when Cloze is entered as a single predictor in an rERP analysis, it reveals to be an important predictor, while at the same time the residuals reveal that this model in fact poorly fits the data. Crucially, the invocation of LMER underlines an advantage of harnessing the rERP layer as a

machine learning layer: If one is not interested in inferential statistics within the individual models, standard linear regression models can be fitted within each subject and time point separately, meaning that there is an optimal, analytical solution to each regression problem. Indeed, this eschews the need for any iterative parameter estimation procedures as are typically required to fit linear mixed effects models, and thereby any model convergence problems.

The machine learning perspective also speaks to the issue of partial collinearity, the situation in which two or more predictors are non-identical yet correlated, thereby potentially affecting the degree to which we can interpret the resulting regression coefficients. As Smith and Kutas (2015a) point out, the presence of collinear predictors does not violate any assumptions underlying least squares regression, and hence of the rERP framework. Moreover, while they do discuss how the effects of collinearity can be examined (using variance inflation factors, VIFs), they recommend to enter all relevant predictors, and to then look for rERPs of interest to the scientific question at hand. Crucially, as we are typically not interested in inferential statistics on the individual coefficients, but rather employ the rERP framework as a machine learning layer, the worst case effect of collinearity is that the influence of individual predictors cannot be isolated, thus yielding under informative rERP analyses. We therefore also recommend to start with a maximal rERP analysis, including all relevant predictors of interest. One can then study the waveforms and coefficients, and see how the analysis changes when one or more predictors are removed. Here, one could also consider the generalizability of the fitted models; that is, while we fitted our models on all trials within a given subject, electrode, and time point, another approach would be to fit them on only a subset of the data, and to then assess their generalization to unseen data. This could for instance inform on whether the models are overfitting the data, as well as on whether an analysis transfers to data from other experiments. Indeed, even if an analysis minimizes residuals on one data set, this does not mean it necessarily generalizes to other data as well (within or across studies). Moreover, even on seen data, non-zero coefficients may not always reflect a true effect of the factors instantiated by the corresponding predictors, for instance due to data sparseness or noise. In such cases, to further validate an rERP analysis, and/or to see how an analysis transfers to unseen data, generalization performance should be taken into account (also see the literature on temporal response functions—which are highly similar to rERPs—in which model cross-validation is common practice; e.g., Crosse, Di Liberto, Bednar, & Lalor, 2016).

To demostrate the utility of the rERP framework as as a viable tool to explain WCS-derived inconsistencies, we have incrementally derived an rERP analysis of the results reported in a recent study on language comprehension by Delogu et al. (2019, DBC). DBC manipulated plausibility and association across three conditions, and found that relative to a baseline condition, in which the target word was both plausible and strongly associated, decreasing plausibility only led to an increase in P600 amplitude in the 600–1,000 ms time-window, while decreasing both plausibility and association led to an increase in N400 amplitude in the 300–500 ms time-window, which continued into a sustained negativity. Indeed, these WCS-derived results appear internally inconsistent with respect to the effect of plausibility: decreasing plausibility seems to increase P600 amplitude in one condition, but relatively less so in another. Our rERP analysis, however, which explicitly models the LCS underlying the observed WCS, revealed that this apparent inconsistency can be explained by means of a quantitative combination of plausibility and association. That is, the regression coefficients revealed that while decreased plausibility pulls the waveforms to be more positive, starting from 400 ms onwards until the end of the epoch, decreased association simultaneously pulls the waveforms to be more negative from 300 ms onwards until the end of the epoch. Crucially, as the absolute coefficients for association are larger than those for plausibility, association is the stronger force. This explains the observed WCS, and in particular the relatively weaker increase in positivity for the condition in which both plausibility and association are decreased; that is, the opposite pull of plausibility and association, of which the latter is stronger than the former, leads to a net negativity. Finally, the relative strength of the coefficients does not only vary in time, but also in space, explaining the fronto-central gradient that is apparent in both conditions with decreased plausibility: more posterior sites go more positive.

The rERP analysis further revealed that when association is kept constant across trials, a significant late positive deflection—a P600 effect—emerges for both conditions in the 600–1,000 ms time-window. Conversely, keeping plausibility constant across trials yields a significant negativity in the 300–500 ms time-window—an N400 effect—for the combined manipulation of plausibility and association, which sustains until the end of the epoch. Indeed, in terms of neurophysiology, this supports the mapping that association is reflected in the N400 component, while plausibility is indexed by the P600 component. Furthermore, it suggests that the sustained negativity arises because these components overlap from 400 ms onwards, and that their summation leads to a net negativity in the 600–1,000 ms P600 time-window. Finally, the relative strength of these components seems to vary across scalp sites, with the N400 being more pronounced at frontal sites, and the P600 more prominent at posterior sites. Taking together, the WCS-derived inconsistencies in the DBC results thus seem to arise from spatiotemporal component overlap between the N400 and the P600.

Crucially, the observation of spatiotemporal overlap between the N400 and the P600 implies that the generators

underlying these components are active simultaneously in time for the larger part of the ERP epoch, and that they combine in the scalp-recorded signal. Many of the apparent WCS-derived inconsistencies, however, common within and across studies, derive from the assumption that the N400 does strictly precede the P600 in time, hence taking these components to be spatiotemporally independent. Indeed, we believe this assumption to be incorrect, and that these inconsistencies can in fact be explained by factoring in spatiotemporal overlap between the N400 and the P600. This requires a paradigmatic shift in which the focus of investigation moves from WCS to LCS, which has implications for experimental design, statistical analysis of results, and neurocognitive theorizing.

Experimental designs should carefully identify which relevant continuous (and/or categorical) predictors derive neatly from the experimentally manipulated factors, can be adequately quantified through stimulus pre-testing, and be segregated from each other if they contribute to the signal at the same point in space and time (see above). Second, analysis of results should focus on how the factors that are manipulated in a given design combine in space and time to yield the observed signal, and critically, how each of these factors independently modulates the signal when the other factors are held constant. The rERP framework extends traditional approaches to ERP analysis with the tools to investigate this: It allows for modeling LCS in terms of how relevant (continuous and categorical) regression predictors combine in space in time, and crucially, it allows for isolating factors by keeping others constant. The core challenge of the rERP analysis is to arrive at regression models that minimize the difference between observed and estimated voltages. Ideally, however, this process should be facilitated by careful experimental design.

Finally, neurocognitive theories which attribute specific computational operations to ERP components should factor in that the processes indexed by these components may overlap in time, for instance by positing architectures in which these processes are organized in a temporally cascaded manner or in which they dynamically interact. An example of the latter is the Retrieval–Integration account of the electrophysiology of language processing (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Brouwer et al., 2012; Brouwer & Hoeks, 2013). On this account the N400 component indexes retrieval of word meaning, while the P600 component indexes integration of this word meaning into the unfolding utterance representation. Crucially, rather than assuming that Retrieval strictly precedes Integration, the account assumes that each word involves a reverberating cycle in which Retrieval and Integration processes dynamically interact. Indeed, Retrieval–Integration theory predicts spatiotemporal overlap between the N400 and P600, and can account for the DBC results (see Delogu et al., 2019, for discussion). Moreover, assuming the functional-neuroanatomic mapping of Retrieval/N400 onto the left posterior middle temporal gyrus (lpMTG) and Integration/P600 onto the left inferior frontal gyrus (lIFG) brouwer2013time, the Event-Related Optical Signal (EROS) results by Tse, Squires, and Hillyard (2007) support such reverberating—and thus overlapping—processing dynamics.

In sum, we believe that apparent inconsistencies in ERP results may be reconciled by shifting focus from WCS to its underlying LCS. More generally, we have demonstrated how the rERP framework, which extends traditional approaches to ERP analysis with an additional layer of statistical estimation, offers a powerful means to explicitly model and investigate LCS. While the framework can be applied to existing data, we believe that careful consideration of future designs may aid data analysis by carefully identifying what continuous predictors are relevant, how stimulus pre-tests may adequately quantify these predictors on a by-item basis, and how different predictors may be segregated if they combine in the signal at any given point in time. Finally, neurocognitive theories should take into account that the processes they attribute to different ERP components may be organized in a cascaded manner and potentially interact dynamically.

ACKNOWLEDGEMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose.

ETHICS STATEMENT

No new human data were collected for the present work. This work models the data collected by Delogu et al. (2019). Their study was conducted in accordance with the ethics approval granted by the Deutsche Gesellschaft für Sprache (DGfS) to the last author in written form. All participants gave written informed consent and were paid for taking part in the experiment.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.14961

DATA AVAILABILITY STATEMENT

Standalone code and the data required to replicate this analysis is available at: https://github.com/hbrouwer/dbc2019rerps.

ORCID

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

REFERENCES

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on 'semantic P600' effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73.

Brouwer, H., & Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Frontiers in Psychology*, *8*, 1327.

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

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

Brouwer, H., & Hoeks, J. C. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, *7*, 758.

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

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.

Donchin, E., Ritter, W., & McCallum, W. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. H. Koslow (Eds.), *Event-related brain potentials in man* (pp. 349–411). New York, NY: Academic Press.

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., … Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883–1908.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, *15*(6), 883–899.

Handy, T. C. (Eds.). (2005). *Event-related potentials: A methods handbook*. Cambridge, MA: MIT Press.

Heikel, E., Sassenhagen, J., & Fiebach, C. J. (2018). Time-generalized multivariate analysis of EEG responses reveals a cascading architecture of semantic mismatch processing. *Brain and Language*, *184*, 43–53.

Hoeks, J. C., 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.

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.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, *85*(1), 1–36.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperberg, G. R., Kreher, D. A., Sitnikova, T., Caplan, D. N., & Holcomb, P. J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language*, *100*(3), 223–237.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Luck, S. J. (2005a). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luck, S. J. (2005b). Ten simple rules for designing ERP experiments. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook*. Cambridge, MA: MIT Press.

Luck, S. J., & Kappenman, E. S. (Eds.). (2012). *The Oxford handbook of event-related potential components*. Oxford: Oxford University Press.

Näätänen, R. (1982). Processing negativity: An evoked-potential reflection. *Psychological Bulletin*, *92*(3), 605.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*(4), 375–425.

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Matthew Husband, E. (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*, *375*(1791), 20180522.

Rugg, M. D., & Coles, M. G. H. (1995). *Electrophysiology of mind: Event-related brain potentials and cognition*. Oxford: Oxford University Press.

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

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

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*(4), 387–401.

Tse, C.-Y., Lee, C.-L., Sullivan, J., Garnsey, S. M., Dell, G. S., Fabiani, M., & Gratton, G. (2007). Imaging cortical dynamics of language processing with the event-related optical signal. *Proceedings of the National Academy of Sciences*, *104*(43), 17157–17162.

van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190.

How to cite this article: Brouwer H, Delogu F, Crocker MW. Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *Eur. J. Neurosci.*2021;53:974–995. https://doi.org/10.1111/ejn.14961