Modelling Non-linear Relationships in ERP Data Using Mixed-effects Regression with R Examples

Antoine Tremblay
Dalhousie University, Halifax, Nova Scotia, Canada

Aaron J. Newman
Dalhousie University, Halifax, Nova Scotia, Canada

Author Note

Corresponding author: Antoine Tremblay, NeuroCognitive Imaging Laboratory, Dalhousie University, Halifax, NS B3H 4R2, Canada. Tel.: (902) 494-1911. A.T. was supported during data collection by a Social Sciences and Humanities Research Council of Canada (SSHRC) doctoral fellowship while in the Department of Linguistics at the University of Alberta, Edmonton, Canada (2007–2009). A.T. was supported during analysis and writing by a SSHRC post-doctoral fellowship in the Department of Psychology and Neuroscience at Dalhousie University, Halifax, Canada (2011–2013). A.J.N. was supported by a SSHRC Standard Research Grant.

Abstract

In the analysis of psychological and psychophysiological data, the relationship between two variables is often assumed to be a straight line. This may be due to the prevalence of the general linear model in data analysis in these fields, which makes this assumption implicitly. However, there are many problems for which this assumption does not hold. In this paper, we show that in the analysis of event-related potential (ERP) data, the assumption of linearity comes at a cost and may significantly affect the inferences drawn from the data. We demonstrate why the assumption of linearity should be relaxed and how to model nonlinear relationships between ERP amplitudes and predictor variables within the familiar framework of generalized linear models, using regression splines and mixed-effects modelling.

This paper has been written in LaTeX using `Sweave` and `R`, and the source document is provided as supplementary material. The data, a `pdf` of this paper, the `.Rnw` file used to write it, and the `R` code used to generate all of the analyses, tables, and figures presented here are available in package `nlEEG`. The package is available as supplementary material for this article, and also from `http://hdl.handle.net/10222/22146`. An additional set of supplementary material entitled *What's Under the Hood: A Brief Introduction to GAMM*, is available from the same URL. These resources can be used to work through the examples, and potentially act as a starting point for the reader's own forays into mixed-effects modelling.

Keywords: EEG; ERP; nonlinear relationships; GAMM; regression splines.

Modelling Non-linear Relationships in ERP Data Using Mixed-effects Regression with R Examples

Introduction

Event-related potential (ERP) datasets are commonly collected in cognitive neuroscience experiments because they offer rich spatio-temporal information about brain activity during perception, cognition, and action. The potential power of ERP data comes with the cost that the datasets tend to be large and quite complex. Thus the richness of the data demands analytical techniques that are appropriate to fully describe both the complexity of the data, and control for factors that might interfere with such analyses. In trying to deal with such complex data, it is very common to assume that the relationship between ERP amplitudes/latencies and an independent variable is linear. For example, in a study that investigates the relationship between age and amplitude of the N1 ERP component, one would assume that a given increase in age is associated with a consistent increase (or decrease) in the amplitude and/or latency of the N1, regardless of whether the increase is from 23 to 24 years old or from 69 to 70 years old.

In reality, there are many problems for which the answer is not a straight line (Rupert, Wand, & Carroll, 2003; Pinheiro & Bates, 2000; Wood, 2006; H. Wu & Zhang, 2006; Hastie & Tibshirani, 1990; Baayen, Kuperman, & Bertram, 2010; Tremblay & Baayen, 2010; Kryuchkova, Tucker, Wurm, & Baayen, 2012) and ERPs are no exception. For example, Carbon, Schweinberger, Kaufmann, and Leder (2005) investigated the effect of "Thatcherized" faces (in which the eyes and mouth regions are turned upside-down) on N170 amplitudes (a negative component occurring between 130 and 200 ms at occipito-temporal scalp sites). Pictures of normal and Thatcherized faces were presented at either 0, 90, or 180° rotation; only when presented upright are these perceived as severely distorted. Carbon et al. (2005) found a nonlinear effect of rotation on N170 amplitudes in response to faces: Upright Thatcherized faces elicited a smaller N170 than Thatcherized faces rotated 90 or 180°. The effect on N170 amplitudes of the latter two orientations,

however, did not differ. Boutheina, Coutya, Langer, and Roy (2009) also investigated the effects of picture rotation on N170 amplitudes. They used pictures of normal faces that were rotated 0, 22.5, 45, 67.5, 90, 112.5, 157.5, or 180° and found a relatively complex nonlinear effect on N170 amplitudes and latencies (see panel B of Figure 3 in Boutheina et al., 2009). Numerous other examples of nonlinear effects on ERPs can be found in the literature, in areas such as development (Webb, Long, & Nelson, 2005), rhythm perception (Pablos Martin, Deltenre, Rossion, Hoonhorst, & Colin, 2007), and auditory processing (Inui et al., 2010).

The studies mentioned above did not assume linearity, which is a step forward. However, in keeping with the traditional ANOVA approach to statistical analysis, they tested for nonlinear relationships using the common practice of discretizing the variables of interest. That is, they have reduced inherently continuous measures, such as rotation angle (or stimulus duration, interstimulus interval, magnitude, frequency, etc.) to factor variables with only a few levels. Such a practice is likely attributable to the natural development of cognitive neuroscience from behavioural approaches to studying cognition that traditionally rely heavily on ANOVA models. However, it comes with problems of its own. Indeed, not only does discretization induce bias into the analysis, but it also decreases statistical power and increases the probability of finding spurious associations (Cohen, 1983; McCallum, Zhang, Preacher, & Rucker, 2002). Fortunately, such factorization is neither desirable, nor necessary to detect nonlinear relationships.[1]

[1]Although we advocate here that continuous variables should not be discretized, there are circumstances in which such variables should be divided into categories. For example if the values of a variable happen to be distributed in such a way that there are two distinct peaks separated by a large gap, then it would make sense to dichotomize it. It goes without saying that an inherently binomial variable should be entered in a model as a factor variable.

Relaxing the assumption of linearity: A set of simple examples

Assuming linearity when it does not hold may lead one (i) to over-look important structure in the data, (ii) to conclude that there is no relationship between two variables when in fact there is one, and (iii) to under-estimate the strength of the relationship between a dependent and an independent variable (Sahai & Ageel, 1997). To illustrate these points, we simulated four data sets (based on the `gamSim` function from `R` package `mgcv`; Wood, 2012; R Development Core Team, 2013) and compared models that assume linearity to ones that do not (for the sake of this example, we will gloss over how the assumption of linearity was relaxed for the latter analyses). In Figure 1, the panels on the left show the models that were fitted assuming linearity, whereas the ones on the right show the models that were fitted without making this assumption. In the first simulated data set, shown in the top panels (A and B), the dependent (Y) variable does not correlate with the predictor variable, and neither of the two models report that it does. The 2 point difference in Akaike's Information Criterion (AIC; Akaike, 1973) indicates that both models are equivalent.[2]

Panels C and D show data in which there is a linear relationship between the response and predictor variables, which is captured by both linear and nonlinear models. Indeed, the nonlinear model fit results in a straight line since there is no nonlinearity in the relationship, and the AIC values indicates that both models are just as likely given the data. The simulated relationship between the dependent and independent variables in

[2]AIC is a relative measure that indexes how likely a model is given the data (relative to other models). It is calculated as *2 × the number of parameters in the model - 2 × the log likelihood of the model* (i.e., more complex models will tend to have higer AIC values). Statistical models can be compared by way of AIC, where those with smaller values reflect more likely models. By convention, we consider that two models differ if the difference in AIC value is equal to or greater than 5, meaning that the model with the lower AIC is 12 times more likely given the data than the one with the higher AIC. We refer the interested reader to Symonds and Moussalli (2011) and Burnham, Anderson, and Huyvaert (2011) for more details about AIC and model selection.

panels E and F is quadratic. The model for which linearity is assumed (panel E) not only fails to capture the quadratic relationship, but it also falsely concludes that there is no significant relationship between X and Y. The model shown in panel F, however, is able to capture the quadratic trend and comparison of the AIC values for the two models confirms that this is a better fit. In panels G and H, the simulated relationship is more complex, taking on a "wavy" function. Although the model for which linearity is assumed does find that there is a significant relationship between X and Y, the actual nature of the relationship is only part of the underlying truth. This leads the model to under-estimate the strength of the relationship between X and Y, and the accuracy of the prediction varies as a function of X. Conversely, the model for which linearity is not assumed, illustrated in panel H, is able to capture the true nature of the relationship, and correspondingly has a much lower AIC value.

Modelling Non-linear Relationships

Testing for nonlinear relationships can be seen as a natural extension of testing for linear relationships. Linear regression is familiar to virtually anyone with a modicum of training in statistics, and allows one to test for a linear relationship between two variables as shown in the left-hand columns of Figure 1. Thus for example in the analysis of the effects of image rotation in the Carbon et al. (2005) study described above, the levels of rotation could have been treated as values along a continuum of possible rotation angles, rather than as categorical levels. The problem with this approach, as demonstrated in the examples in Figure 1, is that if the relationship is nonlinear then a linear fit will not be optimal. In contrast, pairwise testing of different "levels" of rotation is able to detect which (if any) pairs of levels differ. Certain relationships can also be detected using a set of linear contrast weights, as in Helmert or polynomial contrasts. However, this approach does not generalize easily to many levels of a factor, and because each pairwise or contrast test assumes independence correction for multiple comparisons should be applied — potentially

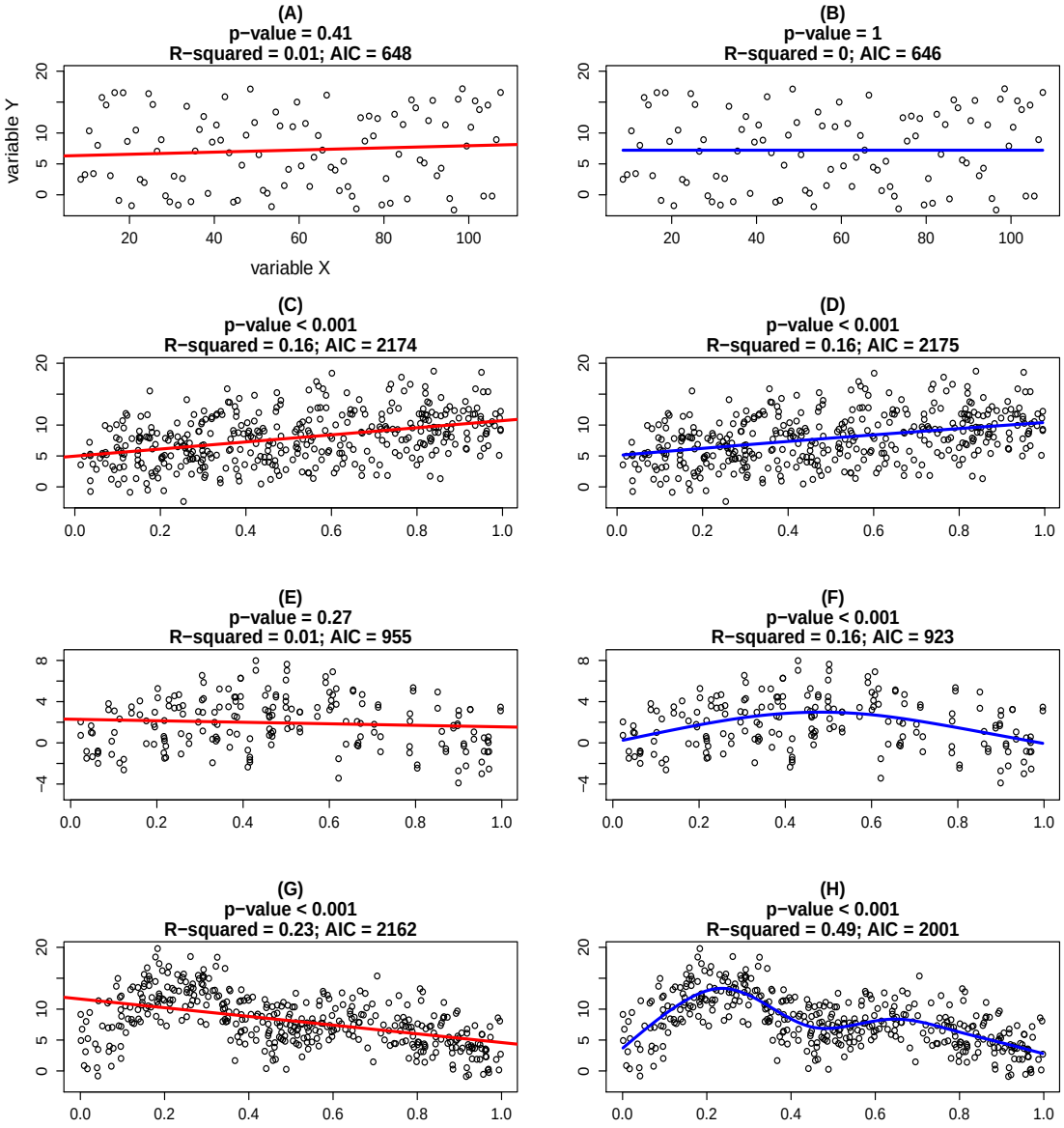

Figure 1. Assuming linearity (left column; red lines) versus not assuming linearity (right column; blue lines).

weakening sensitivity to extant effects.

Non-linear relationships can be modelled without discretization by applying a series of linear transformations to the predictor variables. Regression splines, which are thought to be the ideal series of transformations and the most computationally efficient (Wood, 2006, p. 148), are curves made up of sections of polynomials joined together at points

termed "knots" so that they are continuous in value, as well as first and second derivatives. The specific spline equation used in the model fitting algorithm we use here (i.e., generalized additive mixed-effects regression; see below) is from Gu (2002):

$$R(x,k) = \frac{[(k-\frac{1}{2})^2 - \frac{1}{12}][(x-\frac{1}{2})^2 - \frac{1}{12}]}{4} - \frac{[(|x-k|-\frac{1}{2})^4) - \frac{1}{2}(|x-k|-\frac{1}{2})^2 + \frac{7}{240}]}{24} \quad (1)$$

where x is a vector of covariate values and k is a knot location.

The spline function merely serves to transform variables. The actual modelling of the data will be performed with generalized additive mixed-effects modelling, which we briefly describe below.

Generalized Additive Mixed Effects Modelling

Generalized additive mixed-effects modelling (GAMM) is a relatively recent development in statistics (e.g., Hastie & Tibshirani, 1990; Gu & Wahba, 1991; Wood, 2006) and has recently been applied to ERP data in several published papers, including Tremblay (2009), Tremblay and Baayen (2010), and Kryuchkova et al. (2012).[3] Mixed-effect modelling (e.g., GAMM) is a natural tool for modelling repeated measures (Lin & Zhang, 1999; H. Wu & Zhang, 2006; L. Wu, 2010), including repeated measurements in ERP data (Bagiella, Sloan, & Heitjan, 2000; Vossen, van Breukelen, Hermens, van Os, & Lousberg, 2011). This type of model captures the dependency between repeated measurements – such as for example within subjects as well as within stimuli – by modelling the variance-covariance matrix of the error term. Details about mixed-effects modelling, as well as its potential advantages over repeated measures ANOVA, can be found in a number of recent papers and books (Bagiella et al., 2000; Pinheiro & Bates, 2000; Faraway, 2005;

[3]Generalized additive (mixed-effect) modelling is a widely accepted data analysis method that has been used in well over a thousand papers from a wide variety of domains of inquiry. See section *Domains of Inquiry in which GAMM has been Used* in supplementary material *What's Under the Hood: A Brief Introduction to GAMM*.

H. Wu & Zhang, 2006; Gelman & Hill, 2007; Baayen, 2008; Baayen, Davidson, & Bates, 2008; Quené & van den Bergh, 2008; Zuur, Ieno, Walker, Saveliev, & Smith, 2009; L. Wu, 2010; Vossen et al., 2011).

A GAMM is a non-parametric regression model with the general additive structure:

$$g(\mu_{ij}) = \mathbf{X_i^*}\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots + \epsilon_{ij} + \mathbf{Z_{ij}}\alpha_{\mathbf{i}}, \quad (2)$$

where $\mu_{ij} \equiv \mathbb{E}(Y_{ij})$, Y_{ij} is a response variable (e.g., amplitude), $\mathbf{X_i^*}$ is a row in the matrix for a strictly parametric model component, θ is the corresponding parameter vector, the $f(x)$'s are smooth functions of the covariates, ϵ_{ij} are $N(0, \sigma^2)$ measurement errors, and $\mathbf{Z_{ij}}\alpha_{\mathbf{i}}$ are $N(0, D(\sigma^2))$ random effects (Hastie & Tibshirani, 1990; Faraway, 2005; Wood, 2006; Keele, 2008). The basis we use here for our smooth functions, $f(x)$, are the regression splines defined in Equation (1).

Like ANOVA, GAMM also fits within the generalized linear mixed effects model (GLMM). ANOVA and GAMM differ in that (1) GAMM can model more complex random-effects structures (e.g., crossed, independent random effects; Baayen et al., 2008; Quené & van den Bergh, 2008); (2) GAMM is robust to violations of sphericity if the correct random effect structure is used (Bagiella et al., 2000; Baayen et al., 2008; Vossen et al., 2011), eliminating the need to correct for this post hoc using methods that are known to be either overly conservative (Greenhouse-Geisser) or liberal (Huynh-Feldt), (3) GAMM enables one to appropriately model imbalanced data (Bagiella et al., 2000; Gelman & Hill, 2007; Baayen et al., 2008; Vossen et al., 2011), a common situation in ERP data, and (4) the model fitting objective is augmented by a "wiggliness" penalty enabling one to model non-linear relationships that do not over- or under-fit the data in a manner that reduces subjectivity and circularity (Wood, 2006; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

To fully make use of these capabilities, data should not be averaged across trials or any other variable. Rather, a GAMM should be fitted on the un-averaged ("single trial")

data. Beyond allowing robust, accurate estimates of variance, this practice has the added benefit of allowing one to model nonlinear relationships between dependent and independent variables in multi-dimensional space.

Striking a Balance Between Model Fit and Model Smoothness. By specifying the model in terms of smooth functions rather than detailed parametric relationships it is possible to avoid otherwise cumbersome and unwieldy models. The "wiggliness" of a smooth function (i.e., the trade off between over- and under-fitting the data) is controlled by the integrated square of the second derivative of a covariate, $\int_0^1 [f''(x)]^2 dx$, multiplied by λ, a smoothing parameter (Wood, 2006). As λ tends towards infinity, the estimate of $f(x)$ will become a straight line, whereas $\lambda = 0$ will result in an un-penalized (wiggly) estimate. One way of (objectively) determining λ is through generalized cross validation (GCV). As Hastie and Tibshirani (1990) put it, "it is a way to *let the data show us the appropriate functional form*" of a smooth (p. 1; emphasis is theirs). The GCV objective is to minimize the error between the model-predicted value of a missing data point and its real value. In brief, the flexibility of a GAMM is determined by λ and GCV rather than by the number and location of spline knots (Wood, 2006).

Number of Knots and Knot Locations. By default, knot are located at a covariate's quantiles. Given that model smoothness is determined by λ and GCV, the exact number of knots to use is not generally critical. Nevertheless, it should be large enough that one is reasonably sure of having enough degrees of freedom to represent the underlying "truth" reasonably well, but small enough to maintain reasonable computational efficiency (Wood, 2006, also see the help page for function *choose.k* in R).

Probability Values for Smooth Terms. To calculate a smooth's probability value, a Wald statistic, T is first computed:

$$T = f' V_f^* f \tag{3}$$

where f is the vector of values of a smooth term, and V_f^* is the rank r pseudo-inverse of the

corresponding Bayesian covariance matrix V_f, and r is the estimated degrees of freedom (*edf*) for the term (Wood, 2013). T is then compared to a χ^2 distribution with degrees of freedom equal to *edf* for the term plus α (e.g., 0.05) times the number of estimated smoothing parameters.

Confidence Intervals. Smooths have Bayesian confidence intervals around them, which are obtained by taking the quantiles from the posterior distribution of the $f(x)$'s (Marra & Wood, 2012).

Goals of the Present Study

The goal of this paper is to demonstrate how relaxing the assumption of linearity can lead to better modelling of ERP data. We also demonstrate how to model nonlinear relationships using GAMM. This paper has been written in LaTeX using `Sweave` and `R`, and the source document is provided as supplementary material. The data, a `pdf` of this paper, the `.Rnw` file used to write it, and the `R` code used to generate all of the analyses, tables, and figures presented in this paper are available in package `nlEEG`. The package is available as supplementary material or from `http://hdl.handle.net/10222/22146`.[4] The paper also has a second set of supplementary material entitled *What's Under the Hood: A Brief Introduction to GAMM* also available from `http://hdl.handle.net/10222/22146`. These resources can be used to work through the examples, and potentially act as a starting point for the reader's own forays into GAMM analysis. This paper can, however, be read and understood without viewing the accompanying `R` code.

Example ERP Data

The ERP data used here was collected in the context of an immediate free recall experiment described in Tremblay (2009) and Tremblay and Baayen (2010). The goal of this experiment was to investigate whether frequently used four-word sequences, such as

[4]The `R` code chunks contained in this paper are also available as individual `.R` files in folder *nlEEG/vignettes/Rcode_chunks*.

end of the year, *I don't really know*, and *at the same time*, may be (de)composed via the application of compositional rules or stored and retrieved as wholes (in which case they should show effects of the frequency of co-occurrence of the four-word sequence). This was achieved by examining whether the frequency of such phrases affected early ERP components such as the anterior N1 (N1a) and the posterior P1. Both of these components peak approximately 110–150 ms after stimulus presentation, and were previously found to be modulated in studies of single words by linguistic variables including frequency and length (Sereno, Rayner, & Posner, 1998; Assadollahi & Pulvermüller, 2003; Hauk & Pulvermüller, 2004; Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Murphy, Roodenrys, & Fox, 2006; Penolazzi, Hauk, & Pulvermüller, 2007). Tremblay (2009) reported that higher-frequency four-word sequences indeed elicited more negative N1 and less positive P1 amplitudes than low-frequency sequences.

However, additional data were collected in this study that were not analyzed in Tremblay (2009) or Tremblay and Baayen (2010). In the present paper, we investigate whether working memory capacity, the length of the second word of a sequence, and the position in a list where a four-word sequence was presented (i.e., whether it was presented first, second, ..., fifth, or sixth in a block), and their interaction affected ERPs in the memory task.[5]

In the analyses presented here we focused on the N1a as the ERP component of interest. The N1a is known to be sensitive to spatial attention as well as lexical frequency, probability of occurrence, and word length (e.g., Luck, 2005; Murphy et al., 2006; Penolazzi et al., 2007, and references cited therein). For instance, high-frequency words elicit greater N1a amplitudes than low-frequency words, and longer words elicit larger N1a amplitudes than shorter ones. Because in this paper our focus is on demonstrating nonlinear analysis

[5]The second word of a sequence appeared, more often than not, right where the fixation cross was presented (centre of the screen). Thus, the second word of a sequence was the first word that participants saw. It stands to reason that the longer the second word of a sequence, the greater the amplitude of the N1 should be given that longer words elicit larger anterior negativities (Murphy et al., 2006).

and comparing it to a linear approach, we limited our analyses to a single electrode (Fz, located along the midline, midway between the nose and the vertex of the head) where the N1a component was maximal, and we do not discuss the implications of our results with respect to the cognitive or linguistic literature. Nevertheless, we do provide the R code to perform a GAMM analysis of the whole scalp, as well as results, in section *GAMM Analysis of Entire Scalp in 80–180 ms Window* in supplementary material *What's Under the Hood: A Brief Introduction to GAMM*.

Participants

Ten right-handed female students from the University of Alberta were paid for their participation in the experiment. (Mean age = 23.4; *SD* = 1.6; min/max = 22/27). All were healthy native speakers of English, had normal or corrected-to-normal vision, and did not report any neurological deficits. Participants had a mean handedness score (Oldfield, 1971) of 79.5/100 (*SD* = 15.8). We assessed participants' reading span and working memory capacity (henceforth WMC) using an adaptation of the Daneman and Carpenter (1980) test presented on a PC using E-Prime (Mean WMC score = 73/100; *SD* = 10.4).

Study design

The stimulus list consisted of 432 four-word sequences taken from the *British National Corpus* (Davies, 2004). Some examples are *end of the year*, *I don't really know*, *at the same time*, *I have to say*, *it would be a*, *at the age of*, *this is not a*, *we've got to get*, and *I think it's the*. Frequencies, which were obtained from the *Variations in English Words and Phrases* search engine (Fletcher, 2008), ranged from 0.03 to 105 occurrences per million. The stimuli were divided into 72 blocks. Each block was divided into 18 trials, where in each trial six four-word sequences were randomly presented one at a time in the middle of the screen with an inter-stimulus interval of roughly 4000 ms. Sequences subtended on average $\sim 5^\circ$ x 0.4° visual angle; the longest four-word string (*becoming increasingly clear that*) subtended $\sim 8^\circ$ x 0.4° visual angle. At the end of each trial (i.e.,

after having seen six four-word sequences), participants were prompted to type in as many sequences as they could recall.

EEG Recordings and Processing

Electroencephalogram (EEG) recordings were made with active Ag/AgCl electrodes from 32 locations according to the international 10/20 system (`www.biosemi.com/headcap.htm`) at the midline (Fz, Cz, Pz, Oz) and left and right hemisphere (Fp1, Fp2, AF3, AF4, F3, F4, F7, F8, FC1, FC2, FC5, FC6, C3, C4, T7, T8, CP1, CP2, CP5, CP6, P3, P4, P7, P8, PO3, PO4, O1, O2), as well as the right and left mastoids, and referenced online at the common mode sense active electrode. Electrodes were mounted on a nylon cap. Eye movements were monitored by electrodes placed above and below the left eye and at the outer canthi of both eyes, which were bipolarized off-line to yield vertical and horizontal electro-oculograms (EOG). Analog signals were sampled at 8,192 Hz using a BioSemi (Amsterdam, The Netherlands) Active II digital 24 bits amplification system with an active input range of ± 262 mV per bit and were band-pass filtered between 0.01 and 100 Hz. Note that Biosemi uses active electrodes and therefore can tolerate high scalp impedances.[6]

The digitized EEG was initially processed off-line using `Analyzer` version 1.05 (Brain Products GmbH, Gilching, Germany): It was re-referenced to the average of the right and left mastoids, downsampled to 128 Hz (using an `Analyzer` script), band-pass filtered from 0.01 to 30 Hz using a forward-backward filter combination where each of the filters was comprised of a two pole zero phase infinite impulse response Butterworth filter, and corrected for eye movements and eye blinks by regressing out the vertical and horizontal EOGs (Gratton, Coles, & Donchin, 1983). The processed signal was then segmented into

[6] As mentioned in `http://www.biosemi.nl/forum/viewtopic.php?t=486&highlight=impedance`, for active electrodes "EEG currents do not (significantly) flow via the electrodes because the input impedance of all current biopotential amplifiers is very high. So, typical electrode impedances (smaller than a few hundred kOhm) do not influence measured EEG voltages". Our impedances were around 25 Ohms.

epochs of 3000 ms (1500 ms before and after stimulus onset). Each epoch was baseline corrected on the 1500 ms segment immediately preceding stimulus onset.

An `mp4` movie of the scalp topographies through time can be downloaded from `http://hdl.handle.net/10222/22146`. Figure 2 is an animation of the scalp topography through time. The very first frame of the animation depicts the scalp topography at time $t = 141$ ms where the N1a was maximal. It is apparent from the scalp animation as well as from the bottom panel of Figure 2, that a negative deflection around electrode Fz — the N1a — began at $t \sim 78$ ms, peaked at time $t = 141$ ms ($\sim -5\mu V$), and returned to baseline at time $t \sim 180$ ms. Our window for analysis was thus chosen as 80–180 ms.

Results

The three predictor variables (fixed effects) under investigation were (1) the position of the four-word sequence in the set of six items presented in a block (*Position*), (2) the length of the second word in the four-word chunk (*Length*)[7], and (3) the working memory capacity of the participant (WMC). These predictor variables were transformed prior to plotting or analysis. We first mean-centered *Position*, *Length*, and WMC by subtracting the mean of a variable from each of its values. The main purpose of this centering was to increase numerical stability and remove any spuriously high correlation that may arise between random intercepts and random slopes as well as between fixed and random variables (Hofman & Gavin, 1998; Kreft & De Leeuw, 1998; Baayen, 2008). We first present the results from the behavioural analysis and then move on to the ERP analysis.

Behavioural Results

We investigated whether *Length*, *Position*, WMC, and their interactions affected the probability of correctly recalling a four-word sequence. In order to be correctly recalled, a four-word sequence had to be recalled exactly. That is, if the target sequence was *in the*

[7]The second word was chosen because it most often overlapped the center of the monitor where subjects fixated between trials

middle of, any response other than *in the middle of* was considered to be incorrect (e.g., *in the middle*, *in the middle and*, *in the middle of a*, or *at the middle of*). However, we accepted minor misspellings such as *in the mdle of* or *n the midle of*.

Table 1

Result of the behavioral analysis. Pos = Position. Lngth = Length, WMC = Working Memory Capacity. edf = estimated degrees of freedom.

	edf	Chi.sq	p-value
Pos	4.6	367	< 0.001
Lngth	7.8	23.3	0.004
WMC	1	4	0.045
Pos:Lngth	5.5	11.3	0.136
Pos:WMC	8.4	30.8	0.001
Lngth:WMC	1	2.5	0.112
Pos:Lngth:WMC	9.5	14.3	0.325

The optimal GAMM model is summarized in Table 1. The model included significant $Position \times Length$ and $Position \times WMC$ interactions. Note that although the summary table states that p-value for the $Position \times Length$ interaction was greater than 0.05, it can be concluded from both the AIC comparison and the 95% confidence interval that this interaction should remain in the model.[8] Results of the behavioural analysis are depicted in Figure 3. In panel (A), it can be seen that *Length* had a small effect on the probability of correctly recalling a four-word sequence. More specifically, recall probabilities were greatest when the second word was 2 to 6 letters long at positions 5 and 6. Panel (B) illustrates the $Position \times WMC$ interaction, which exhibited an overall recency effect whereby more recently presented four-word sequences were more likely to be correctly recalled. The

[8]We do not consider the 3-way $Position \times Length \times WMC$ or the 2-way $Length \times WMC$ interactions given that $p > 0.05$ *and* they were not warranted by way of AIC comparisons.

recency effect possibly reflected an advantage in activation strength for the sequences that were presented more recently within a block (Jones & Oberauer, 2013). However, it was more pronounced for participants with higher working memory capacity scores. Furthermore, people with a WMC score below 0.7 did not reliably recall four-word sequences presented in positions 1, 2, and 3, whereas people with a WMC score above 0.7 exhibit a small primacy effect in position 1.

ERP Results

Discretized Variables and ANOVA

Although our main goal in this paper is to demonstrate the application of nonlinear analysis using continuous variables, we first show the data treated categorically. *Position* was treated as having 6 levels while *Length* and WMC were dichotomized as high/low based on a median split. Figure 4 shows waveforms and mean N1a amplitudes for each level of each of these three variables. No effect of *Length* is apparent in this figure. However, there do appear to be amplitude differences for different levels of *Position* (middle panel) — in particular, items that appeared first in the lists show a more negative N1a than subsequent positions. As such, there may be a nonlinear relationship between *Position* and N1a amplitude. The bottom panel also suggests the possibility of an effect of WMC, where participants with a higher WMC showed a more negative N1a than those with a lower working memory capacity.

Figure 5 shows waveforms and mean amplitudes for two- and three-way combinations of levels of the independent variables (corresponding to the two- and three-way interactions in an ANOVA model). There is a possible *Length* $\times$ *Position* interaction, where long words in first position may have elicited a more negative N1a than short words in first position. There also may be a *Length* $\times$ *WMC* interaction and a *WMC* $\times$ *Position* interaction. In the former case, shorter words may have elicited a smaller N1a in participants with a low working memory capacity. In the latter case, the position effect on N1a amplitudes may

have been stronger in participants with high than low working memory capacity. Finally, there is possibly a three-way interaction (bottom panel), where long words in the first position elicited a more negative N1a in participants with low than high working memory capacity (the leftmost black and red bars). We performed a repeated measures ANOVA to test these observations. It revealed that only the main effect of *Position* was reliable ($F(5,96) = 6.2$, $p < 0.001$).

GAMM Analysis Assuming Linearity

The present analysis (using only the parametric component of GAMM) simply treated each variable as continuous. Results are shown in Table 2. This model suggests that there was a significant main effect of *Position* and a significant $Length \times WMC$ interaction. These effects are depicted in Figure 6.

The *Position* effect shown in panel (A) suggests that the amplitude of the N1a decreased (became more positive) as position increased.[9] It is illustrated in panel (B) that the length of the second word of a sequence modulated the amplitude of the N1a. More specifically, it increased (became more negative) as the working memory capacity score of a participant increased for words 1 to 6 letters long, but decreased (became more positive) for words 7 to 12 letters long.

While it is notable that these effects were found under the assumption of linearity, Figure 4 suggested that *Position* had a nonlinear relationship with N1a amplitude. Describing the source of this main effect would necessitate drawing inferences over the six levels of Position from a relatively large number of pairwise comparisons.[10] Fitting a single nonlinear function to *Position* could simplify the description and interpretation of the

[9]Because the N1a has a negative polarity, we refer to less negative amplitudes as "decreases" and more negative amplitudes as "increases".

[10]While contrasts such as Helmert could also describe this relationship somewhat more parsimoniously, we had no a priori assumption as to the shape of the relationship between Position and N1a amplitude so the choice of appropriate contrast weights would have to be made post hoc.

effect without recourse to cumbersome post hoc testing. In addition, it may be the case that the assumption of linearity missed important structure in the data. In the next section, we explore models for which linearity is *not* assumed.

Table 2

*Model for which linearity **was** assumed. Pos = Position. Lngth = Length, WMC = Working Memory Capacity.*

	Estimate	*Std.Error*	$t-value$	$p-value$
Pos	0.142	0.018	8.0	<0.001
Lngth	−0.037	0.029	−1.3	0.198
WMC	−0.645	1.376	−0.5	0.639
Pos:Lngth	0.005	0.010	0.5	0.635
Pos:WMC	−0.255	0.167	−1.5	0.125
Lngth:WMC	0.472	0.152	3.1	0.002
Pos:Lngth:WMC	−0.034	0.095	−0.4	0.722

Moving Away from the Assumption of Linearity

The possibility of nonlinear relationships between our predictor variables and N1a amplitude is supported by Figure 7, in which mean N1a amplitude at each sampled level of each variable are plotted, along with a form of best-fitting nonlinear functions (lowess smooths).[11] The *Length* effect (top left panel) appears to be curved with relatively constant (and small) N1a amplitudes from values 1 to roughly 6, after which N1a amplitude increases with increasing word length. The effect of *Position* (bottom left panel) shows an opposite pattern. The amplitude of the N1a is most negative in the first position, less so in second, third, and fourth positions, and then increases slightly in the fifth and

[11] A lowess smooth uses locally-weighted polynomial regression. 'Local' is defined as a window of a certain span around each data point (e.g., 50 data points to the left and to the right). Each data point within a window is influenced (i.e., weighted) by its "neighbours". See `?lowess` in `R` for more details.

sixth positions. The WMC effect (top right panel) is more complex, with an initial increase in N1a amplitude from the lowest to low-middle WMC levels, followed by a decrease and then a subsequent increase for people with the highest WMC.

We thus fitted a GAMM without assuming linearity. It included a three-way interaction $Position \times Length \times WMC$ and all lower order interactions and main effects, with by-subject and by-item random effects. We allowed the model to use up to 227 degrees of freedom, but GAMM deemed that 55 degrees of freedom were needed to appropriately model the data. This model was more likely given the data than the one assuming linearity (model assuming linearity: $df = 9$, $AIC = 357382$; model not assuming linearity: $df = 55$; $AIC = 357228$; difference: $df = 46$, $AIC = 154$). Results are provided in Table 3. The plot of the three-way interaction is shown in Figure 8.

Table 3

Model for which linearity was ***not*** *assumed. Pos = Position. Lngth = Length. WMC = Working Memory Capacity. edf = estimated degrees of freedom.*

	edf	F	$p - value$
Pos	4	27.6	<0.001
Lngth	1	1.0	0.332
WMC	1	0.2	0.629
Pos:Lngth	9	2.9	0.001
Pos:WMC	12	4.5	<0.001
Lngth:WMC	4	2.8	0.016
Pos:Lngth:WMC	24	1.6	0.020

As in the analysis assuming linearity, here we found a significant main effect of *Position*. This confirmed the pattern observed in both Figures 4 (discretized variables) and 7 (continuous variables) whereby N1a amplitudes were largest for words in the first position but were more or less consistent across subsequent positions. The lack of

significant main effects for *Length* or WMC are also consistent with the linear analysis, suggesting that although the nonlinear curves shown in Figure 7 appeared to describe the data more closely than a straight line would have, nevertheless these relationships were not statistically reliable.

More striking differences between the linear and nonlinear analyses reside in the interaction structure of the models. While the linear model only included a significant *Position* $\times$ WMC interaction, the model for which linearity was not assumed contained a significant three-way interaction *Position* $\times$ *Length* $\times$ WMC. Recall that this interaction was predicted above based on our examination of the plots shown in Figure 5. The three-way interaction is depicted in Figure 8, where the *Length* $\times$ *WMC* interaction is plotted at each *Position*. Note that regions shaded in white correspond to portions of a surface where the 95% confidence interval included 0 μV. Given the principle of marginality (Venables, 1998), we did not attempt to interpret the two-way interactions or main effects.

Figure 8 shows a complex relationship between *Position*, *Length*, and WMC. It is apparent that the amplitude of the N1a increased with *Position*, and that the strongest N1a occured at *Position* 6 as can be evidenced by the dark blue areas in the panel for this position. Moreover, the areas shaded in white indicate that the significance of the N1a decreased as *Position* increased, that is, the regions where the 95% confidence interval included 0 μV became larger and larger: Whereas the whole surface was essentially significant at positions 1 and 2, the N1a at positions 4, 5, and 6 was less and less reliable, especially for longer words and for participants with a WMC score below 0.65. Starting at position 4, the amplitude of the N1a elicited by 8–12 letter words was essentially 0 μV while the N1a to 3 letter words became increasingly unreliable. At position 6, only 1–2 and 4–8 letter words exhibited a significant N1a, and only for participants with a WMC score above 0.65.

In the behavioural analysis, we had found that *Position*, *Length*, and WMC affected the probability of a four-word sequence being correctly recalled. More specifically, (i)

sequences for which the second word was longer were less likely to be correctly recalled, (ii) more recently presented four-word sequences were more likely to be correctly recalled, (iii) this effect was more pronounced for people with higher WMC scores, and (iv) people with a lower WMC score correctly recalled four-word sequences only when they were presented in the last three positions. Our ability to relate the ERP results back to behaviour, however, is seriously compromised by the fact that we conflated successfully- and unsuccessfully-recalled sequences. In the next section, we attempt to characterize the relationship between the probability of recall of a four-word sequence and N1a amplitudes. We additionally investigate whether the length of the second word of a sequence, the working memory capacity of a participant, and the position in which a sequence was presented modulated this relationship.

Linking Behaviour and Scalp-recorded Potentials

In order to relate the behavioural and ERP results, we first computed the probability of recall of each four-word sequence for each participant from the model obtained in the behavioural analysis. We subsequently fitted a GAMM that included a four-way interaction *Position* $\times$ *Length* $\times$ $WMC\times$ *Probability of Recall* to N1a amplitudes. Results are provided in Table 4 as well as in Figure 9.

This model was much more likely than the previous one (model with 3-way interaction: $df = 55$, $AIC = 357228$; model with 4-way interaction: $df = 106$, $AIC = 351639$; difference: $df = 51$, $AIC = 5589$). The four-way interaction was significant ($F(40, 53229) = 0.4$, $p < 0.001$) thus suggesting that the N1a was not only affected by *Position*, *Length*, and WMC, but was also modulated by *Probability of Recall*. In Figure 9, each row corresponds to one of the quantiles of the *Probability of Recall* variable (0, 0.16, 0.25, 0.39, and 0.92) whereas each column corresponds to one of five *Length* values (1, 4, 6, 9, and 12). The most noteworthy feature in Figure 9 is the increase of N1a amplitudes with the probability of correctly recalling a four-word sequence. Specifically, it was equal to 0

Table 4

Linking behavioural and ERP data. Pos = Position. Lngth = Length. WMC = Working Memory Capacity. PrRec = Probability of Correct Recall. edf = estimated degrees of freedom.

	edf	F	$p - value$
Pos	4	8.2	<0.001
Lngth	1	2.7	0.100
WMC	1	0.2	0.693
PrRec	3	3.1	0.019
Pos:Lngth	5	3.4	0.002
Pos:WMC	11	2.9	<0.001
Lngth:WMC	1	2.3	0.128
Pos:PrRec	5	2.5	0.019
Lngth:PrRec	1	2.2	0.141
WMC:PrRec	1	1.1	0.317
Pos:Lngth:WMC	12	0.6	0.903
Pos:Lngth:PrRec	3	3.3	0.010
Pos:WMC:PrRec	12	1.2	0.264
Lngth:WMC:PrRec	6	2.2	0.031
Pos:Lngth:WMC:PrRec	40.0	0.4	<0.001

μV when the probability of recall was 0, but increased to -15 μV at position 1 for longer words in people with higher WMC scores when the recall probability was 0.92.

It is known that the N1a is associated with attention, and that attended items typically elicit higher N1a amplitudes (e.g., Luck, 2005; Penolazzi et al., 2007, and references cited therein). If the N1a indexes the amount of attentional resources allocated to encoding a sequence, then it can be concluded that the successful recall of a four-word

sequence is associated with the allocation of greater amounts of attentional resources during the encoding phase. Furthermore, longer sequences presented in the first few positions required greater amounts of attention in order to be correctly recalled, and people with higher WMC scores were able to allocate more attentional resources to these items than people with lower WMC scores.

Discussion

Relaxing the assumption of linearity enabled us to gain a better understanding of the data by capturing important structure that was overlooked in the model for which linearity was assumed. Indeed, only a main effect of *Length* and a $Position \times WMC$ interaction were uncovered in this model whereas the one for which linearity was not assumed revealed a significant three-way interaction $Position \times Length \times WMC$. Moreover, even if these interactions would have reached significance in the model assuming linearity, the association strength between N1a amplitudes and these interactions (as indexed by the percentage of deviance explained, a type of effect size) would have been much lower than in the one for which linearity was not assumed, as can be seen in Table 5. In the model for which linearity was not assumed, the *Position* main effect and the $Length \times WMC$ interaction accounted for roughly twice as much deviance as in the one assuming linearity (i.e., double the effect size). Moreover, the $Position \times Length \times WMC$ in the former model accounted for 337 times more deviance than the same interaction in the latter model (i.e., this effect was 337 times greater). It is important to note that the extra amount of deviance explained by the model for which linearity was not assumed is *not due to data over-fitting*. Indeed, the effects depicted in Figures 8 and 9 were obtained by penalized regression and GCV – a data-driven approach that strikes a balance between under- and over-smoothing the data. If the effects should have been a straight line, GAMM would have penalized them to that extent, as it did in the examples provided in panels (B) and (D) of Figure 1.

By not assuming linearity, we were also able to push our understanding of the data

Table 5

Percentage of the deviance explained by each fixed-effect term of the model assuming linearity and the one not assuming linearity. The terms in bold face correspond to ones that were significant in the model for which linearity was not assumed. Overall, the model for which linearity was assumed accounted for 2.75% of the variance whereas the one for which it was not assumed accounted for 2.88% of the variability in the data (including random effects).

	AssumingLinearity	*NotAssumingLinearity*
Pos	**0.115400**	**0.204800**
Lngth	0.000200	0.000000
WMC	0.000100	0.000200
Pos:Lngth	**0.000200**	**0.073000**
Pos:WMC	**0.003400**	**0.136400**
Lngth:WMC	**0.017600**	**0.035100**
Pos:Lngth:WMC	**0.000400**	**0.134800**

even further by allowing the $Length \times WMC$ smooth surfaces at each *Position* to vary according to the probability of correct recall of a four-word sequence. The results obtained from this model – shown in Figure 9 – were in close agreement with the continuous averages illustrated in Figure 7. More specifically, the N1a (1) increased as *Length* increased, (2) was smaller for people with intermediate WMC scores (for low recall probabilities), and (3) was greater in the first few positions. In the remainder of the discussion, we tackle a few questions that may have arisen while reading the paper.

When Can a Continuous Variable be Discretized. As mentioned in the Introduction, discretizing an inherently continuous variable comes at the cost of decreasing statistical power, increasing the probability of false positives, and increasing the probability of false negatives (Cohen, 1983; McCallum et al., 2002). If for some reason or another a

researcher is interested in dichotomizing a variable, this should be done on the results of an analysis where continuous variables where used and linearity was not assumed. A threshold for category membership can then be derived from the model predicted values.

Hypothesis Testing and Exploratory Analyses. It is often believed that not assuming linearity implies that one is conducting an exploratory analysis. Although it may be true in some cases, an analysis in which linearity is not assumed may just as well be based on *a priori* hypotheses. Both types of analyses can be performed with the same analysis pipeline used here. That is, a hypothesis can be deemed true or false by virtue of the effect of interest being present or not in the most likely model.

Overfitting. It is the belief of some researchers that once the assumption of linearity is relaxed, the risk of fitting idiosyncratic features of the data is greatly enhanced. Given that GAMM uses GCV to strike a balance between over- and under-fitting the data, the fitting of idiosyncratic features is unlikely to arise (see section Generalized Cross Validation in supplementary material *What's Under the Hood: A Brief Introduction to GAMM* for more details and some examples). An example of this was shown in Figure 1 (i.e., GAMM did not over-fit the data). The analysis pipeline also includes the viewing of a smooth's 95% confidence interval, which further protects against over-fitting. Portions of the smooth for which 0 (the baseline) is included in the confidence interval are considered to be flat (however wiggly they may be).

Sample Size. It is sometimes thought that in order to perform mixed-effects modelling one needs a sufficient number of observations. One should not forget that a repeated-measures ANOVA is also a mixed-effects model, which is often used to model small data sets without any problems. The reason we analyse our data at the single-trial level is not because we need a large sample size to perform mixed-effects modelling, but rather to account for sources of bias that would otherwise be lost during the averaging process and irreversibly affect the averages. For example, it affords us the possibility to account for unequal sample sizes between groups, a situation that commonly arises in ERP

data (some trials and/or channels are usually removed). If we were to aggregate the data as is traditionally done, it would appear to contain an equal number of data points in each cell, but in fact the averages would be biased (i.e., they would have been obtained from unequal sample sizes). Another reason regards the removal of outlier data points. If the data is un-averaged, data points with undue influence can be removed based on model residuals, thus making the results more generalizable. If the data is first averaged, outlier data points will have irreversibly affected the averages. In our experience, the need to average the data has only come about for computational tractability.

Note that an analysis for which linearity is not assumed can always be used whatever the size of the data (and still find significant effects). Some examples of this using real data sets with sample sizes between 56 and 180 are provided in file *chunk25-smallSampleSizeExamples.R* (which can be found in the `.Rnw` file or in folder *nlEEG/vignettes/Rcode_chunks* of supplementary material `nlEEG`).

Model Convergence Problems. Model convergence problems are most likely to arise when (1) the researcher is attempting to fit a model with too many predictor variable relative to the number of data point in the data, (2) there are too many redundant predictor variables (high collinearity), and/or (3) the random effect structure is too complex. In the first case, one should not include in the model more predictor variables than the sample size divided by 15 (e.g., 990 data points / 15 = 66 variables). In the second case, the researcher should deal with the collinearity present in their data in some way or another (see, e.g., Tremblay & Tucker, 2011). In the third case, the researcher should fit his or her model with simpler and simpler random effect structures until the model converges. The most important random effects to include in a model are crossed by-subject and by-item random intercepts, which in our experience are always warranted, as well as by-subject adjustments for regions of interest (to account for both individual variation in the scalp distribution of the effects as well as individual spatial correlations between levels of regions of interest; Newman, Tremblay, Nichols, Neville, & Ullman,

2012). Note that, in our experience, models including these random effects always converge. In some analyses of ERP data, adding by-item random intercepts, and especially by-item random smooths, results in models that are computationally intractable. This forces us, unfortunately, to omit them from our models. With smaller sample sizes, computational tractability becomes less of a problem and one can usually include these random effects.

Conclusion

In order to deal with the complexity of ERP data, it is very common to assume that the relationship between ERP amplitudes/latencies and independent variables is linear. This assumption does not always hold, however. Although in some ERP studies the assumption of linearity has been relaxed, these studies have typically employed the common practice of discretizing continuous variables, which has been shown elsewhere to be suboptimal.

In this paper, we have demonstrated how nonlinearities in ERP data can be modeled using generalized mixed-effects modelling in `R`. This revealed that the assumption of linearity may lead one to draw incorrect inferences from his/her ERP data. More specifically, we have illustrated how the assumption of linearity may (i) increase the probability of false negatives, (ii) if a relationship is obtained, provide a limited or incorrect representation of the true relationship between variables, and (iii) under-estimate the strength of the association between variables.

In spite of the clear potential advantage of nonlinear analyses in capturing the true structure of the data, the greatest limitation of this technique is probably the added complexity in interpreting the results. In the present case we have demonstrated that this complexity is not necessarily uninterpretable however, and indeed the results provide a much richer understanding of the data than could be provided by linear assumptions.

On a closing note, whether or not a researcher predicts that the relationship between a response variable and an independent variable is linear, if the independent variable is

continuous and has an adequate distribution we believe that linearity *should never be assumed*. If the relationship is in fact linear, GAMM will conclude that it is linear (as was illustrated in Figure 1, panel D) and results will be that much more convincing. If the relationship emerges as non-linear then theory can be adjusted.

Figure 2. Scalp topography at time $t = 141$ ms where the N1a was maximal. Blue colors indicate negative amplitudes while red colors indicate positive amplitudes. Green indicates an amplitude of 0 μV. The bottom panel shows the average waveform of each of the 32 electrodes overlaid on top of each other (each electrodes is graphed with a different color). The x-axis represents time in milliseconds and the y-axis is amplitude in μV. A vertical black bar marks the peak of the N1a component. An animation depicting the evolution of the scalp topography from -100 to 600 ms can be viewed with `Acrobat Reader`. To play the animation, click on one of the "triangle buttons" right below the figure. The triangle pointing toward the left will play the animation backward. The one pointing to the right will play it going forward. The animation can be manipulated by clicking on the "<" and ">" buttons to move the animation one frame at a time backward or forward, respectively. The "|<" and ">|" buttons make the movie start from the beginning or go to the end, respectively. The "->|<-" button to the right restores the animation to it's default speed. The "-" and "+" buttons to its left and right make the animation go slower or faster, respectively.

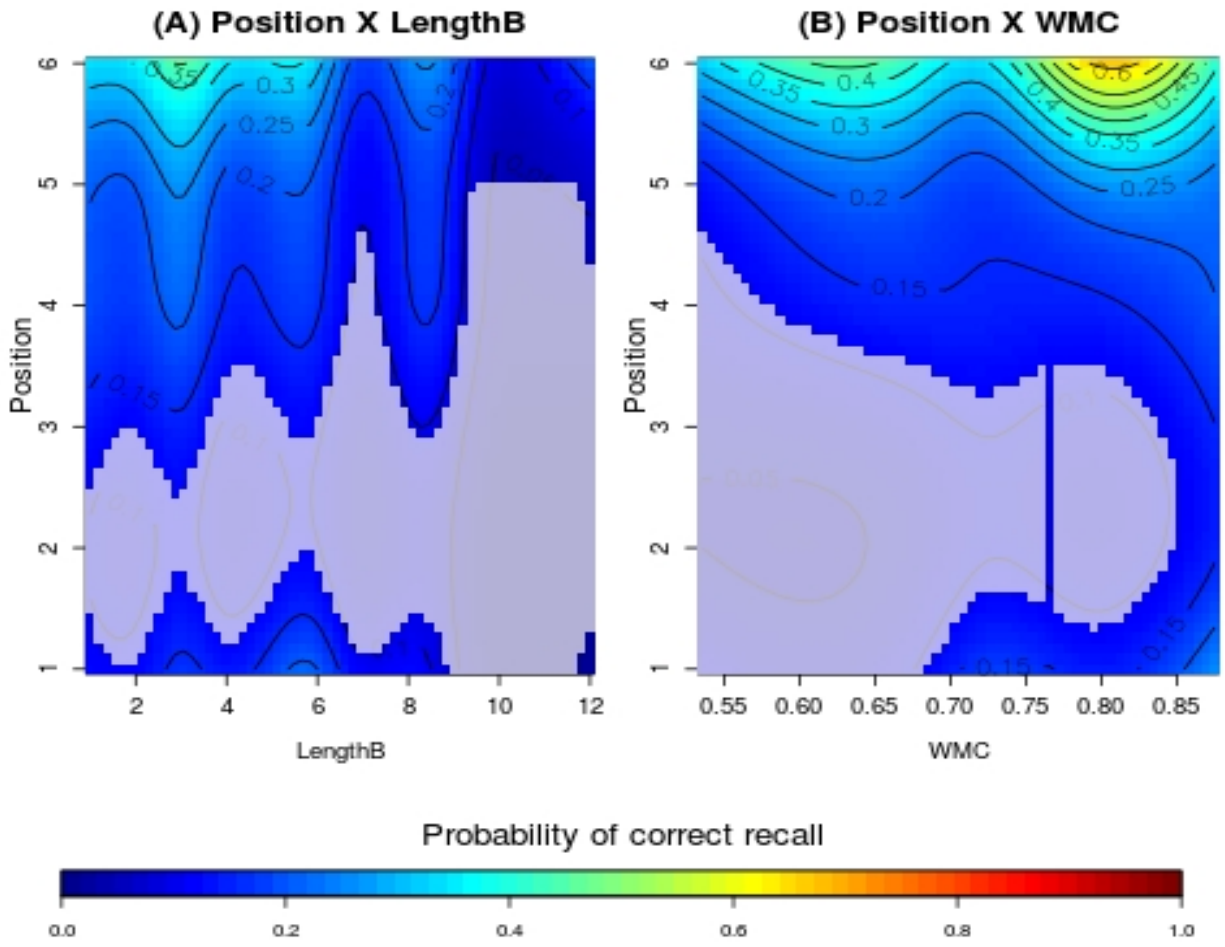

Figure 3. Results of the behavioural analysis. (A): Effect of the *Position* × *Length* interaction on the probability of correctly recalling a four-word sequence. (B): Effect of the *Position* × *WMC* interaction. Bluer colors reflect smaller probabilities of correctly recalling a sequence whereas redder ones indicate higher probabilities. The numbers appearing on the contour lines are recall probabilities. Regions shaded in white included a probability of correct recall of 0 in their 95% confidence interval.

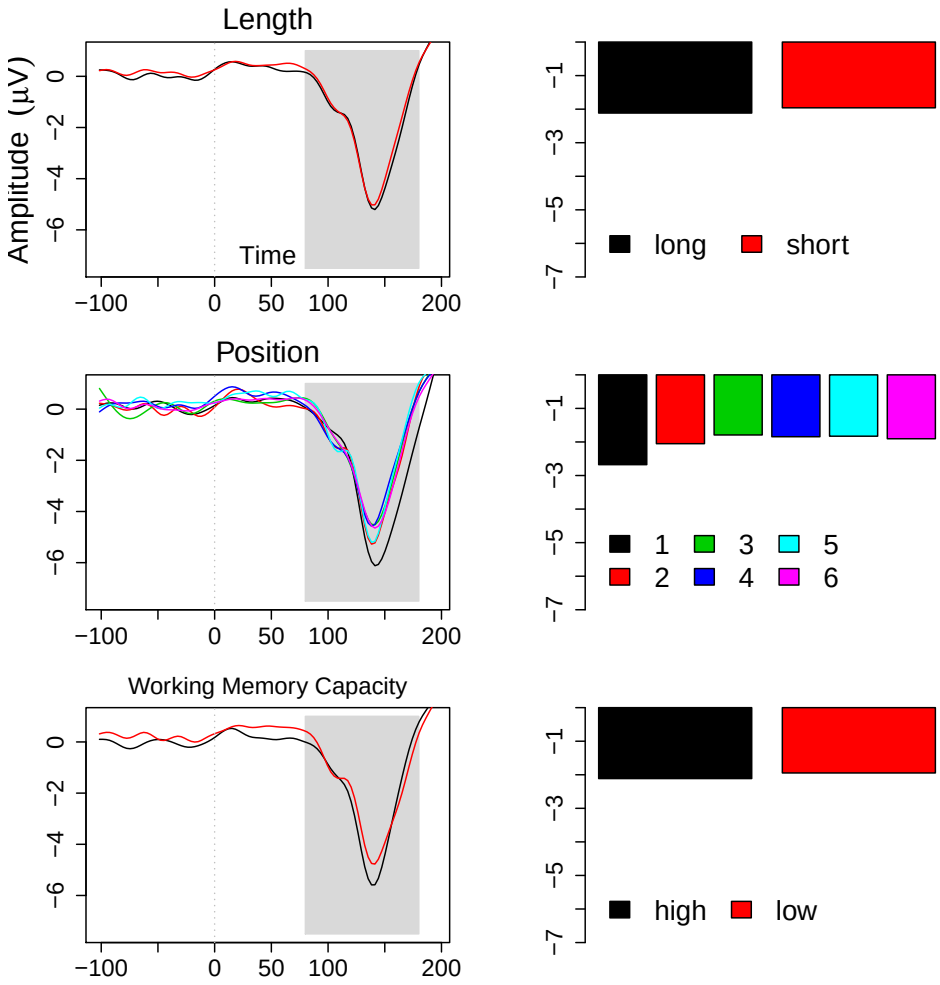

Figure 4. Waveforms (right panels) and mean amplitudes (left panel) of the N1a at electrode Fz, in the 80–180 ms window, plotted separately for discretized levels of each independent variable. The x-axis represents time in milliseconds and the y-axis is amplitude in μV.

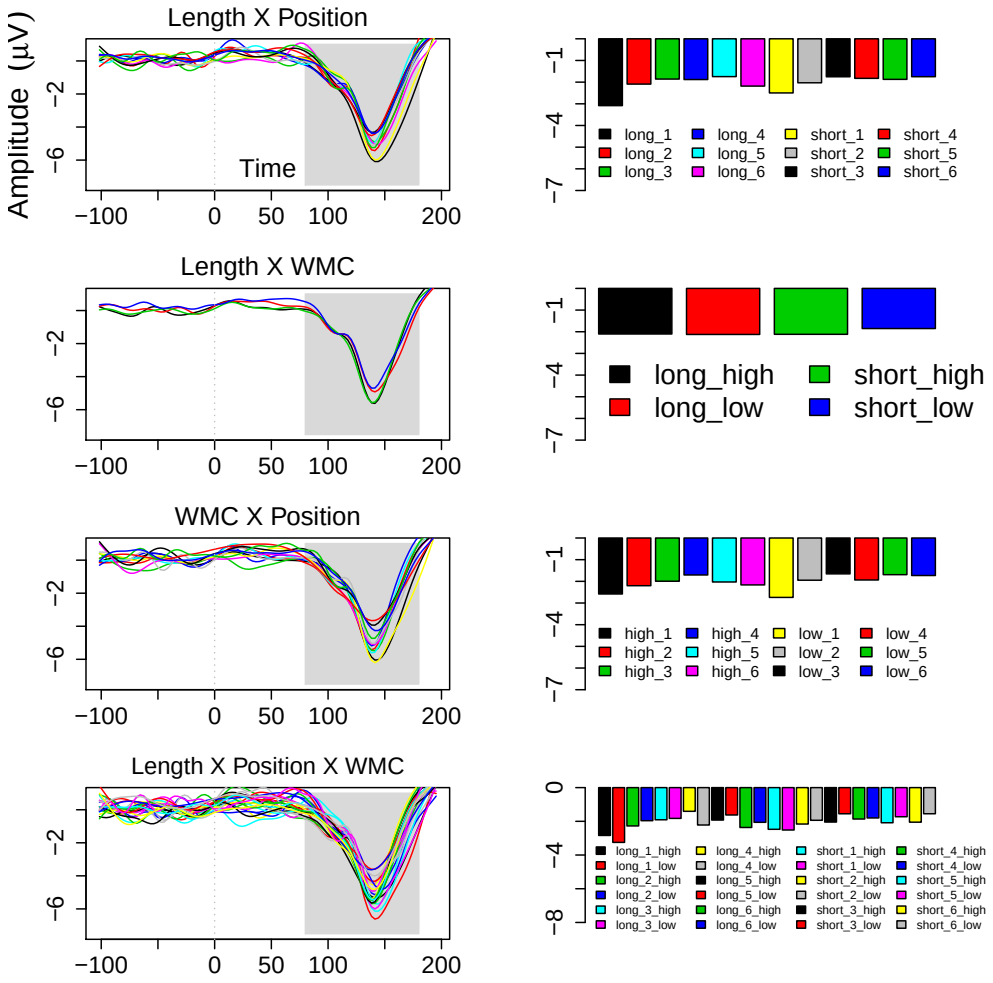

Figure 5. Two-way and three-way interactions between dichotomized levels of *Length*, *Position* and *WMC*. Details are as in Figure 4.

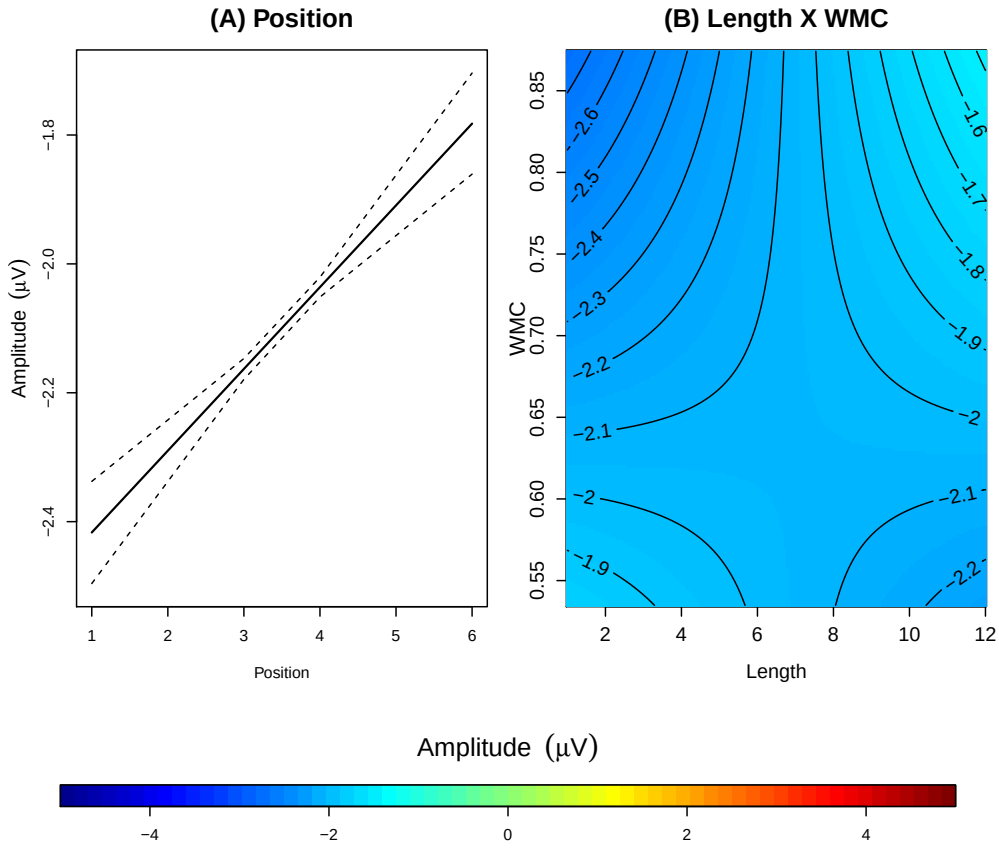

Figure 6. Results of the ERP analysis assuming linearity. In panel (A), the x-axis is *Position* and the y-axis is *Amplitude* (μV). The broken lines are 95% confidence intervals. In panel (B), the x-axis is *Length* and the y-axis is WMC. The amplitude of the N1a is shown using the same color coding as in Figure 2. The scale is provided at the bottom of the figure. The small numbers on the black lines are isovoltage lines with the voltage in microvolts provided.

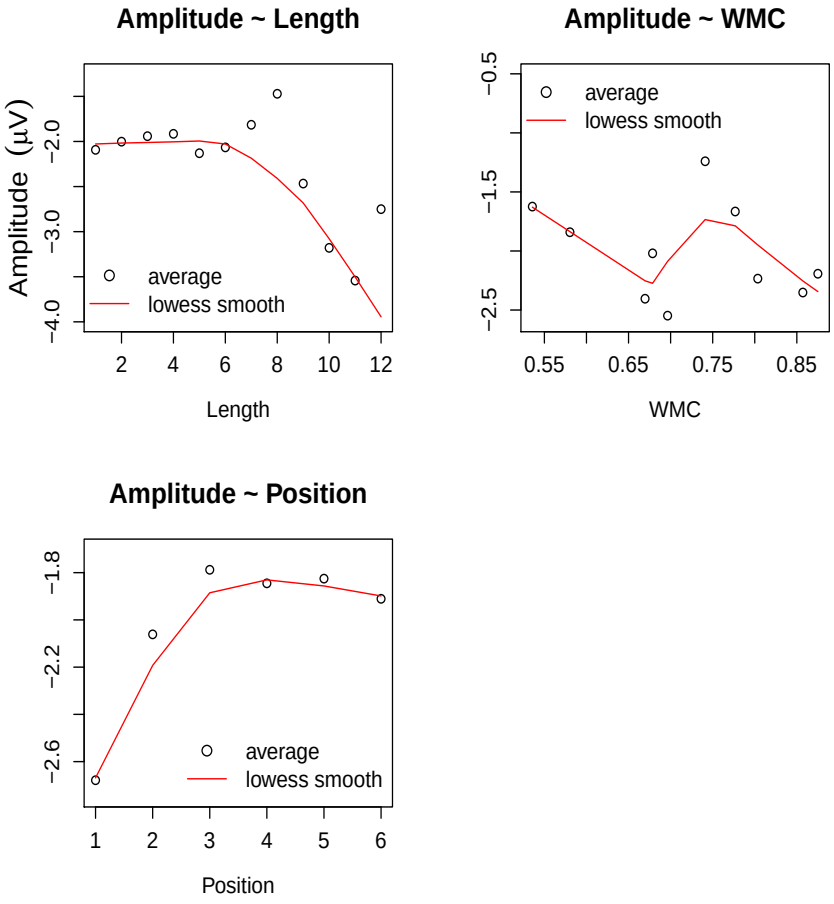

Figure 7. N1a amplitude averages as a function of each one of the predictor variables, treated as continuous (in contrast with the discretized forms of these variables shown in Figure 4). The black circles are the mean amplitude at each measured level of each variable. The red lines are `lowess` smooths of the averages.

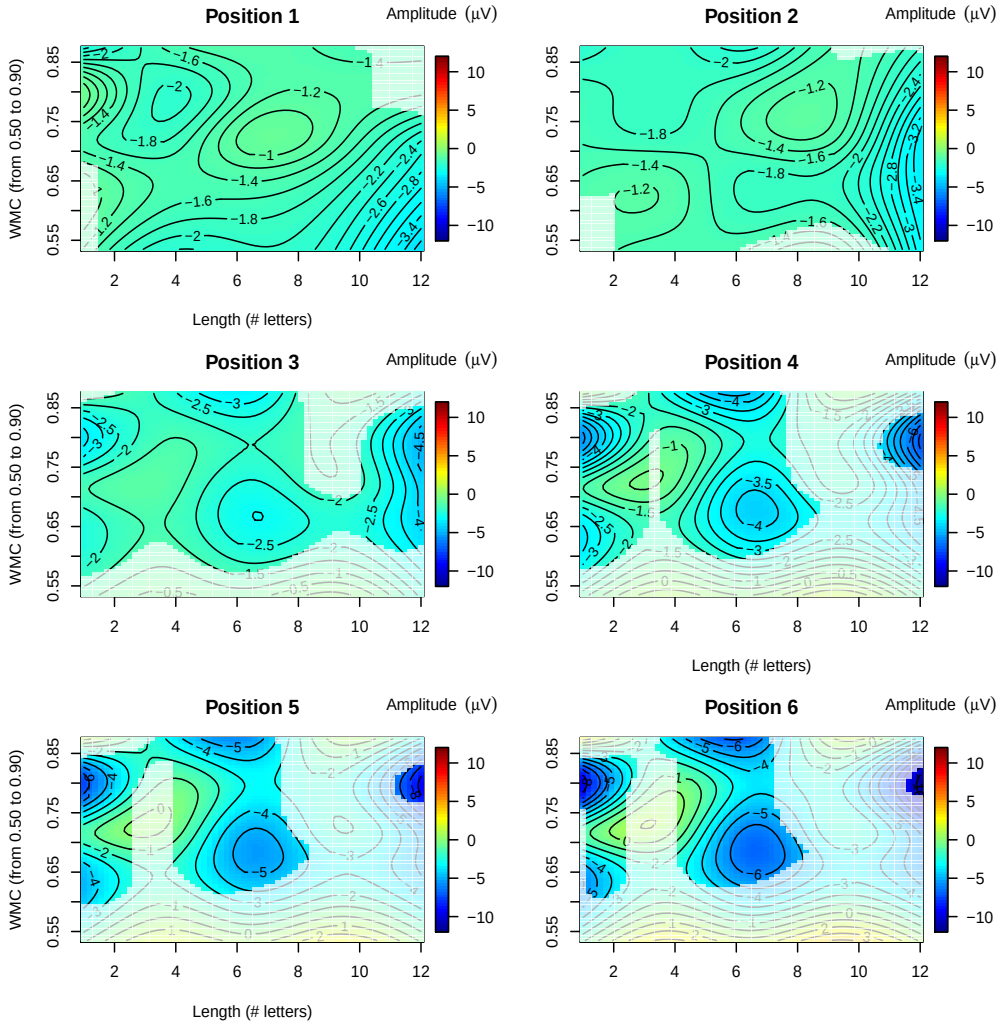

Figure 8. Results of the GAMM analysis were linearity was **not** assumed. Each panel shows the *Length* × *WMC* interaction at each position. The amplitude of the N1a is shown using the same color coding as in Figure 2. The scale is provided at the bottom of the figure. The small numbers on the black lines are isovoltage lines with the voltage in microvolts provided. Regions shaded in white correspond to portions of the smooth that included 0 μV within the surface's 95% confidence interval.

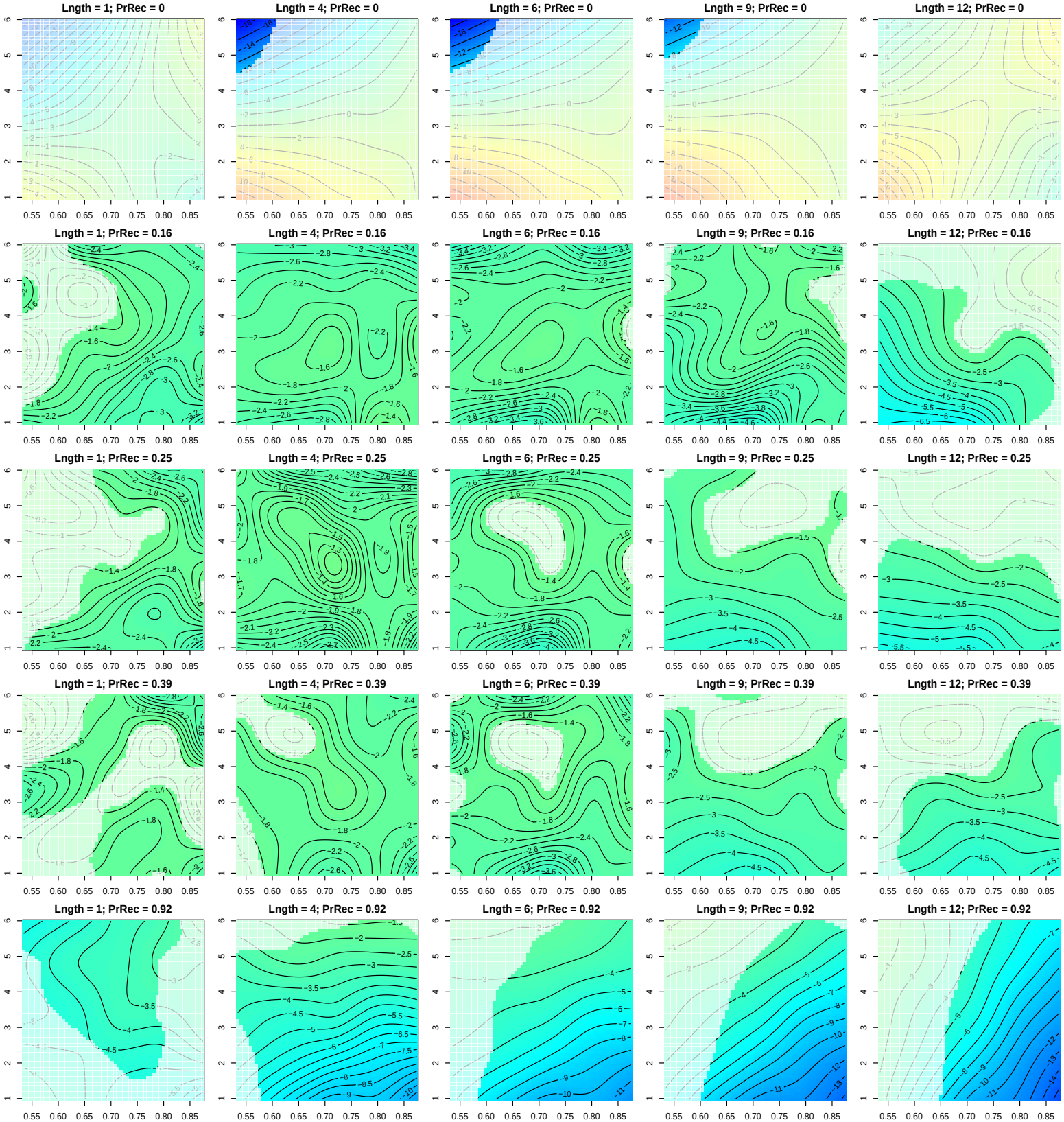

Figure 9. Four-way interaction $Position \times Length \times WMC \times ProbabilityofCorrectRecall$. Each panel shows the $WMC \times Position$ model predicted N1a amplitudes for one of five lengths (columns) and probability of recall values (rows). In each panel, the x- and y-axes are *WMC* and *Position*, respectively. *Lngth* = *Length*. *PrRec* = *Probability of Recall*.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest.

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and frequency: a group study using MEG. *NeuroReport*, *14*, 1183–1187.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using* R. Cambridge, U.K.: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Compounding.* (pp. 257–270). Amsterdam and Philadelphia: John Benjamins.

Bagiella, E., Sloan, R., & Heitjan, D. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*, 13–20.

Boutheina, J., Coutya, J., Langer, C., & Roy, S. (2009). From upright to upside-down presentation: A spatio-temporal ERP study of the parametric effect of rotation on face and house processing. *BMC Neuroscience*, *10*, 1–17.

Burnham, K., Anderson, D., & Huyvaert, K. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*, 25–35. doi: 10.1007/s00265-010-1029-6

Carbon, C.-C., Schweinberger, S. R., Kaufmann, J. M., & Leder, H. (2005). The Thatcher illusion seen by the brain: An event-related brain potentials study. *Cognitive Brain Research*, *24*, 544–555.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–254.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Davies, M. (2004). *BYU-BNC: The British National Corpus.* Available online at http://corpus.byu.edu/bnc.

Faraway, J. J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* New York: Chapman & Hall.

Fletcher, W. H. (2008). *Phrases in english.* Available online at http://pie.usna.edu/index.html.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gratton, G., Coles, M., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr Clin Neurophysiol*, *55*(4), 468–84.

Gu, C. (2002). *Smooth Splines ANOVA Models.* New York: Springer.

Gu, C., & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal of Scientific and Statistical Computing.*, *12*, 383–398. doi: 1369-7412/99/61381

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Model Regression.* New York: Chapman & Hall.

Hauk, O., Davis, M., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*, 1383–1400.

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*, 1090–1103.

Hofman, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*, 623–641.

Inui, K., Urakawa, T., Yamashiro, K., Otsuru, N., Nishihara, M., Takeshima, Y., ... Kakigi, R. (2010). Non-linear laws of echoic memory and auditory change detection

in humans. *BMC Neuroscience*, *11*, 1–13.

Jones, T., & Oberauer, K. (2013). Serial-position effects for items and relations in short-term memory. *Memory*, *21*, 347–365.

Keele, L. (2008). *Semiparametric Regression for the Social Sciences.* New York: Chapman & Hall/CRC.

Kreft, I., & De Leeuw, J. (1998). *Introducing Multilevel Modeling.* London: Sage.

Kriegeskorte, N., Simmons, W., Bellgowan, P., & Baker, C. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*, 535–540. doi: 10.1038/nn.2303

Kryuchkova, T., Tucker, B. V., Wurm, L. H., & Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and Language*, *122*(2), 81–91.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, *61*, 381–400. doi: 1369-7412/99/61381

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

Marra, G., & Wood, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Scandinavian Journal of Statistics*, *39*, 53–74. doi: 10.1111/j.1467-9469.2011.00760.x

McCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Murphy, K., Roodenrys, S., & Fox, A. (2006). Event-related potential correlates of the word length effect in working memory. *Brain Research*, *1112*, 179–190.

Newman, A., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of english. *Journal of Cognitive Neuroscience*, *25*, 1205–1223.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologica*, *9*, 97–113.

Pablos Martin, X., Deltenre, P., Rossion, B., Hoonhorst, I., & Colin, C. (2007). Perceptual biases for rhythm: The Mismatch Negativity latency indexes the privileged status of binary vs non-binary interval ratios. *Clinical Neurophysiology*, *118*, 2709–2715.

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, *72*, 373–388.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus.* New York: Springer.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

R Development Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Rupert, D., Wand, M., & Carroll, R. (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

Sahai, H., & Ageel, M. I. (1997). *Analysis of Variance: Fixed, Random and Mixed Models.* Boston: Birkhauser.

Sereno, S., Rayner, K., & Posner, M. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *NeuroReport*, *9*(10), 2195–2200.

Symonds, M., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*, 13–21. doi: 10.1007/s00265-010-1037-6

Tremblay, A. (2009). *Processing advantages of lexical bundles: Evidence from self-paced reading, word ad sentence recall, and free recall with event-related brain potential recordings.* (PhD Dissertation). University of Alberta, Edmonton, Canada.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 151–173). London and New York: Continuum.

Tremblay, A., & Tucker, B. V. (2011). The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302–324.

Venables, W. (1998). *Exegeses on Linear Models.* Retrieved from `http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf`. (Paper presented to the S-PLUS User's Conference, Washington, DC, 8–9th October, 1998)

Vossen, H., van Breukelen, G., Hermens, H., van Os, J., & Lousberg, R. (2011). More potential in statistical analyses of event-related potentials: a mixed regression approach. *International Journal of Methods in Psychiatric Research*, *20*, e56–e68.

Webb, S. J., Long, J. D., & Nelson, C. A. (2005). A longitudinal investigation of visual event-related potentials in the first year of life. *Developmental Science*, *8*, 605–616.

Wood, S. N. (2006). *Generalized Additive Models.* New York: Chapman & Hall/CRC.

Wood, S. N. (2012). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mgcv` (R package version 1.7-22)

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*, 221–228. doi: 1369-7412/99/61381

Wu, H., & Zhang, J. (2006). *Nonparametric regression methods for longitudinal data analysis.* Hoboken, NJ: Wiley.

Wu, L. (2010). *Mixed Effects Models for Complex Data.* New York: Chapman & Hall.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* New York: Springer.