# A COMPARISON OF STATISTICAL METHODS FOR RELATING INDIVIDUAL DIFFERENCES TO EVENT-RELATED POTENTIAL COMPONENTS

by

Sean R. McWhinney

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2018

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Evidence suggests that individual differences in cognition can influence cortical processing of language violations, as differences are associated with neural recruitment. Individual differences have been found to affect two ERP components that are commonly referenced in studies of language processing. These are the N400, a negative-going deflection with a central parietal distribution, which is sensitive to violations of semantic expectation, and the P600, a positive response seen to violations of phrase structure. Despite the fact that measurable individual differences might influence the latency and topography of these responses, individual differences in cognition are rarely considered, and the ideal method to account for them in ERP studies has not been explored.

This research investigates limitations in statistical approaches for investigating the relationship between ERPs and individual differences. These analyses use three techniques. First, model selection processes are evaluated in order to circumvent problems associated with multicollinearity among numerous individual difference measures. This includes selection of the measures evaluated, depiction of interactions, and specification of random effects. Outcomes of user-specified parameters in each area are characterized to identify an ideal model selection process for a typical EEG data set. Second, by relaxing the assumption of linearity in interactions we aimed to characterize important details that may be lost when only identifying linear effects. The question of sample size required to sustain nonlinear interactions was addressed by using simulated manipulations of sample size to evaluate the propensity for over-fitting with polynomial functions. Third, a non-parametric approach was used to characterize both response topographies and individual difference measure effects through data-driven means, avoiding the requirement to specify *a priori* regions of interest or proficiency bins for significance testing.

Appropriate use cases and limitations for each technique are discussed alongside recommendations for implementing them into future investigations of individual differences in language processing. Considerations made during model specification, both in terms of effect inclusion and the complexity of nonlinear interactions, may improve sensitivity to subtle effects. Moreover, combined with data-driven selection of scalp regions or proficiency bins, the reader is presented with a means to overcome a number of limitations in hypothesis testing.

# List of Abbreviations Used

| | |
|---|---|
| AIC | Akaike Information Criterion |
| ANOVA | Analysis of variance |
| DOF | Degrees of freedom |
| DTI | Diffusion tensor imaging |
| EEG | Electroencephalography |
| ERP | Event-related potential |
| GAMM | Generalized additive mixed effects modeling |
| GLM | General linear model |
| ICA | Independent component analysis |
| ID | Individual difference |
| LME | Linear mixed effects modeling |
| LSpan | Listening span |
| OSpan | Operation span |
| ROI | Region of interest |
| SD | Standard deviation |
| SNR | Signal to noise ratio |
| TOAL-3 | Test of Adolescent and Adult Language 3 |
| TOWRE | Test of Word Reading Efficiency |

# Acknowledgments

For his insight and guidance, I would like to thank my supervisor Dr. Aaron Newman. Dr. Newman's contributions to this work are apparent at every level and have provided me with tools that I use on a daily basis. In addition, the efforts and expertise of my dissertation committee members, Dr. Thomas Trappenberg and Dr. Steven Aiken, have been invaluable, and their contributions to the development of this research are deeply appreciated. My understanding of statistical modeling has been greatly shaped by Dr. Antoine Tremblay, as he dedicated himself to the learning of others. His passion is contagious, and his efforts have been crucial to my success. Lastly, having worked closely with Dr. Shaun Boe and Dr. Timothy Bardouille has inspired an attention to detail and resourcefulness that has influenced my approach to this dissertation, and many projects beyond its scope.

This dissertation would only have been possible through the financial support that has been made available. This includes the National Sciences and Engineering Research Council of Canada, who funded three years of my research. Additional support was provided by the Dalhousie University Department of Psychology & Neuroscience, as well as the RADIANT Graduate Student Scholarship.

This research builds on the efforts of numerous lab members and volunteers. A special thanks to all involved, including Ella Dubinsky, Lisa Beck, Allison Brawley-Hogg, and in particular Kaitlyn Tagarelli who performed all cleaning and pre-processing of the data. Lastly, I would like to thank my wife Alexandria Muise-Hennessey, who has experienced firsthand the joys and frustrations that these past years have entailed. Not only has she provided much-needed support, but input and feedback on this work to improve its quality. This research is the culmination of the efforts and influences of all listed here, as well as many who could not be.

# Chapter 1: Introduction

## 1.1. Electroencephalography as a Method of Inquiry

Traditionally, our understanding of cognition has been limited by the fact that inferences must be drawn using outwardly visible behavioral metrics, such as reaction time, or self-reported assessments from participants. Until relatively recently, this has presented a considerable limitation for developing frameworks to describe cognitive processes, as an unknown and potentially infinite number of combinations of internal processes might result in the same outwardly measurable behavior. However, recent decades have seen improvements in non-invasive neuroimaging techniques, which allow for a specific account of neural activity across distributed systems in the brain at a temporal resolution on the order of milliseconds. One such technology, electroencephalography (EEG) has been used to record the summation of electrical potentials at the scalp, which arise from coordinated neural activity at the cortical surface. Characterizing this activity in response to carefully-controlled stimuli has allowed for a thorough description of the brain's reaction to, and processing of, sensory input. It has also allowed for inference into higher-order cognitive processing of this information. Combining the two approaches, it has allowed for delineation of divergent processing streams, which otherwise might not produce any discernable behavioral difference. These benefits have been invaluable to our development of conceptual frameworks in information processing.

A commonly-used technique in EEG is the averaging of event-related potentials (ERPs), which are recorded at the scalp, and are time-locked to the presentation of a stimulus. ERPs are one of the primary methods of inquiry into mental representations of information in the brain. ERPs are recorded in response to repeated encounters with a stimulus, with the intent that

averaging these recordings together will result in a reduction of any signal that is unrelated to the time-locked stimulus (e.g., unrelated cognitive and physiological processes that result in electrical potential at the scalp). Reducing these signals is critical to uncovering those of interest in an experiment. Conversely, characteristics of the response which are systematically related to presentation of the stimulus should be present in each recording and will remain through the averaging process. Higher-amplitude responses should be relatively more pronounced among unrelated "noise" in the signal, requiring fewer presentations of the stimulus, while lower-amplitude responses require additional averaging to mitigate unrelated signals.

Through this type of characterization, responses to two or more stimuli which differ in some definable way (e.g., in some visual or auditory characteristic) can be contrasted with one another to identify aspects of the response which are associated with the unique traits of a stimulus. This allows for inferences to be drawn regarding the cognitive or physiological mechanisms that are associated with processing these specific characteristics, including association with specific neural anatomical substrates, or on a larger scale with identifying processing streams through a series of distinct brain regions. While this can result in a wealth of information when compared with behavioral reactions to a stimulus, the interpretation of ERP responses is rarely straightforward. For example, while one type of stimulus may result in more neural recruitment than another, as might be hypothesized due to a higher-amplitude response, its effect on a downstream system may be inhibitory. Therefore, a larger-amplitude response may ultimately be associated with reduced output, or an effect on some system which is opposite to what was predicted. Moreover, ERPs comprises both base computation of direct sensory neural input and higher-order cognitive processing. The two rarely occur in isolation and can be difficult to discern from one another. However, strictly-controlled experimental design

has allowed for a degree of delineation of the two and has benefited our understanding of information processing in the brain.

As a practical example of how ERPs can be characterized to understand underlying mechanisms of processing, consider the brain's response to a visual stimulus. Time-locking a recorded average scalp potential to the onset of a stimulus reveals a consistent positive voltage, which peaks in amplitude between 80 and 130 ms following presentation of the stimulus (Mangun, 1995). This positive deflection, which is only visible through averaging the response over many presentations of the same stimulus, is termed a *component*. This component is named the P1 (or alternatively P100), for its positive electrical potential and the timing of its peak amplitude following the onset of a stimulus. Interestingly, the amplitude of the P1 can be manipulated through changing the intensity or brightness of a color being presented, suggesting it reflects processing of basic sensory input (Cobb & Dawson, 1960; Hillyard & Munte, 1984). However, the P1 also appears to reflect higher-order cognition as well, as its amplitude is largest when a stimulus is presented in a region of the visual field to which a participant was paying attention (Van Voorhis & Hillyard, 1977). This component demonstrates that careful manipulations of the characteristics of a stimulus can be used to understand how it's processed and conceptualized, but also that this processing is further mediated by internal cognitive processes. The two occur in a way that, once understood, can be separated through experimental control which attempts to change only one dimension at a time (either visual characteristics or participant attention). Changes in a component's presentation can occur in time (i.e., latency), space (scalp distribution), or in its amplitude.

## 1.2. Functional Associations of the N400 & P600

Isolating experimental manipulations that induce changes in the latency, amplitude or topography of a component can be critical to identifying the cognitive functions with which it is associated. In particular, where the changes in these measures are dependent on the degree of a systematic change in the stimulus, such manipulations can provide insight into mental representations of information or cognitive frameworks. The analysis of the P1 described above represents a general framework which can be used to investigate a variety of components and can help us to understand a variety of cognitive functions. In such cases, the components may not be the focal point of research specifically, but rather they can allow us to understand the processes that they are associated with. The relationship between ERP components and these processes will be the focus of this dissertation, specifically in identifying the most suitable techniques to characterize the link between the two.

One component on which intense focused has been placed to identify its associated neural substrates and functional significance is the N400. This component is identified as a negative-going deflection that occurs approximately 300-500 ms following the perception of a word or image that is semantically incongruent with (i.e., unrelated to) its context (Juottonen, Revonsuo, & Lang, 1996; M Kutas & Hillyard, 1980; Marta Kutas & Federmeier, 2011; Newman, Tremblay, Nichols, Neville, & Ullman, 2012; Nobre & McCarthy, 1994). While the N400 appears to reflect the processing of relationships between objects in general, it has primarily been used to understand language processing. Specifically, it has been used to understand how networks of associations between words and objects are formed. Characteristics of the N400 have been found to be impacted by a wide array of experimental manipulations, each narrowing our understanding of what this component represents. For example, considering the two sentences

(1) *a robin is a bird* and (2) *a robin is not a vehicle*, both are true statements which follow a suitable grammatical structure. However, an N400 is elicited by the second noun in each (either the word *bird* or *vehicle*), and its amplitude is largest in response to the word *vehicle*. This is due entirely to the more distant semantic associations between the two nouns in this sentence, as the more unrelated nature of the words *robin* and *vehicle* is responsible for a larger-amplitude N400 (Fischler, Bloom, Childers, Roucos, & Perry, 1983).

Similarly, when presented with a list of words that follows no rules other than a categorical similarity between these words (i.e., there is no sentence structure), an N400 is elicited in response to the last item, and this response is largest in amplitude when it is categorically unrelated to the items before it (Kutas & Van Petten, 1994). Therefore, while the N400 has frequently been studied in the context of language, these findings suggest that it does not specifically operate in the language domain. Instead it appears to reflect a broader network of semantic association. Furthermore, these effects are not limited to written stimuli, as an N400 has been found in response to spoken language, American Sign Language, line drawings, pictures, faces and even language-like pseudo-words which have no inherent meaning (Kutas & Van Petten, 1994). Given the suite of cognitive functions that have been related to the N400, it does not appear to arise from a single, definable operation, but instead represents a generalized accessing and integration of semantic information across input modalities. Nonetheless, the ubiquitous need to identify the relationships between words during reading and conversation makes it a powerful tool to understand language processing.

The N400 has frequently been studied in tandem with another related component, the P600. Similarly to the N400, the P600 is named for its positive electrical potential, and the timing of its peak amplitude, which occurs approximately 600 ms following the onset of a word which introduces a morphosyntactic violation into a sentence. While the two components are sensitive

to different linguistic rulesets, they can both be elicited through experimental design which contrasts violations of different aspects of sentence structure. That is, while the N400 is thought to index the semantic relationship between items (i.e., words or images), the P600 is instead thought to reflect processing of high-level grammatical structure (Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb, 2007; Nakano, Saron, & Swaab, 2010; Pakulak & Neville, 2010). For example, when compared with the sentence (1) *the child threw the toy,* the sentence (2) *the child throw the toy* should elicit a P600 in response to the word *throw* (Kaan & Swaab, 2003). The result is that sentences can be designed with gradations of violations to either expected semantic associations, grammatical rules, or even both in a single sentence, to elicit either type of response. Moreover, this means that in cases where it's unclear whether damage to a sentence's structure is semantic or morphosyntactic in nature, the type of response elicited can offer insight into how this information was processed. This type of experimental design can reveal divergent processing streams where they might not otherwise be obvious.

As in the example above, holding some aspects of sentence structure constant while carefully manipulating others can help to understand the boundaries of information processing between semantic and morphosyntactic domains. For example, while the relationship between the gender of a pronoun and the noun it describes might be considered a semantic association (e.g., *he* and *Jim*), the fact that violations of this agreement elicit a P600 rather than an N400 suggest that the brain processes this as a violation of a structural more so than a semantic rule (L Osterhout & Mobley, 1995). Similarly, damaging the plausibility of a sentence by reversing the direction of action between two nouns (e.g., by changing *the boy rode the bike* to *the bike rode the boy*) does not alter N400 amplitude, but instead elicits a P600 in response to the final word of the sentence. This is because while the semantic association between the nouns is preserved, violating the structure of the preceding sentence by using an implausible terminal word forces

an attempt at reassessment of the sentence's meaning (Kuperberg, 2007). In this way, modifying the structure of a sentence, or the relationship between words in a sentence, can inform conceptual frameworks of information processing.

## 1.3. The Influence of Individual Differences

When discussing the N400 and P600 in the context of language, these processes reflect not only computation of systemic input, but also rely on semantic associations and understanding of a language's rules and limitations that have been previously learned. Individual differences in disparate aspects of language proficiency (e.g., vocabulary, grammatical understanding), or other aspects of cognition which might dictate methods of information processing (executive processing), could therefore be expected to influence the degree to which either of these components are depicted in response to different types of violations. It is important to note that while we are focusing here on language, the components discussed appear to represent processing of associations and rulesets that span multiple domains. As described above, the N400 is not only sensitive to linguistic stimuli, but to mental representations of objects and even pseudo-words with no inherent meaning (Marta Kutas & Van Petten, 1994). Even beyond the scope of the N400 and P600, individual differences are known to affect ERP responses in varied domains. For example, in an investigation of cognitive demand, assessments of individuals' working memory capacity has been related to P300 amplitude in an *n*-back task (Dong, Reder, Yao, Liu, & Chen, 2015), and this relationship has been associated with meaningful differences in task performance. In another area entirely, differences between individuals in the ability to recognize previously-seen faces has been related both to early differences in the topography of the P100, and differences in the subsequent N170 amplitude (Turano, Marzi, & Viggiano, 2016). This pattern is thought to reflect a difference in the

initial perception and following recognition of faces between individuals. Furthermore, Meyer et al. ( 2017) have found that the well-established link between anxiety and error related negativity may be better predicted using individual P300 amplitude, which mediates the two. The increasing association of abstract individual difference measures with ERPs in domains outside language processing highlights the need for a well-defined link between the two.

In some cases, ERPs themselves may be considered as indicators of individual differences, as an alternative to the functions that they have traditionally been associated with (Meyer et al., 2017). That is, at present we propose to characterize individual difference measures such as facets of language proficiency, which have been operationalized through some testing procedure, but which are not inherently tied to any specific anatomy of physiological process. These psychological constructs are then related to a quantifiable physiological reaction (electrical potential recorded at the scalp). However, the associated physiological processes are not necessarily more representative of any internal language processing framework than the individual difference measures themselves. Indeed, ERPs can be a powerful tool to gain insight into those frameworks, and so understanding the relationship between all three (scalp recordings, individual difference measures, and processing frameworks) is necessary. Put another way, ERP component characteristics can themselves be a revealing marker of differences between individuals. The two must be used in conjunction to understand cognition. This dissertation simply focuses on establishing a link between the individual difference measures and ERPs. While this was done in the context of language specifically, it is hoped that findings may be valuable in numerous domains.

With this in mind, the question of relating the latency or topography of ERP components to individual differences is more general than the question of relating the N400 and P600 specifically to measures of language proficiency. However, language makes for an interesting

avenue of investigation into the processing streams indexed by these components. First, language violations can be strictly controlled, and understanding of a language's rules can be evaluated in relatively standardized ways. It is therefore possible that the influence of individual differences discussed here could apply to processing semantic associations and grammatical structure outside of language domains entirely. Second, understanding of a language's rules does not occur on a single dimension of low- to high-proficiency. Instead, individuals can be more adept in certain areas of linguistic rulesets (e.g., grammar, phonology, or sentence comprehension) and vocabulary than others. This makes language a multi-faceted area in which different types of individual proficiencies or even seemingly-unrelated cognitive characteristics might interact with cognitive processing of language-related information in complex ways that reveal more domain-general processing characteristics. Indeed, the question of whether language processing relies on cognitive systems which are specific to language, or whether they are intertwined with more general cognitive faculties, has been long-standing (Christiansen & Chater, 2008; Fedorenko, 2014; Fodor, 1983). Therefore, language is perhaps not the simplest area in which the link between individual differences and measures of cortical processing can be evaluated. However, it is one which can be characterized on multiple dimensions within a single experiment, and which can potentially inform experimental design in numerous areas.

## 1.4. Individual Differences in Language Processing

Mounting evidence suggests that native language proficiency, as measured through grammatical ability and vocabulary size, can impact the latency, amplitude, and topographical distribution of ERP responses during sentence processing (Liang and Chen, 2014; Moreno and Kutas, 2005; Newman et al., 2012; Pakulak and Neville, 2010; Tanner, 2013; Tanner et al., 2014; Tanner and Van Hell, 2014; Weber-Fox et al., 2003). This evidence suggests differential

recruitment of brain regions based on proficiency or other cognitive factors (Bryll, Binder, & Urbanik, 2013). Specifically, EEG has revealed that the cortical response to words that are semantically congruent with a sentence's context are larger in amplitude (more positive) for individuals with a stronger understanding of English grammar and vocabulary size (Newman et al., 2012). Interestingly, however, these measures of proficiency were not related to the response to semantically incongruent words, suggesting that low- and high-proficiency individuals processed the two similarly. As the N400 is typically calculated as the difference in amplitude between the two conditions, this pattern could be described as a larger-amplitude N400 response in individuals with higher proficiency. This pattern was seen both in Native speakers of English, and in participants who learned English later in life.

The N400 is not the only component which has been tied with individual differences in proficiency. Using the same measures of grammar understanding and vocabulary size as a measure of proficiency, participants were shown either well-formed sentences, or sentences which contained an error in the phrase structure (a phrase structure violation). This violation was associated with a larger, and more widely-distributed P600 response (Pakulak & Neville, 2010). As described above, the P600 amplitude is understood to be mostly sensitive to grammatical cues in a sentence, and as such it is not surprising that it should be influenced by proficiency with grammatical structure. However, this pattern is often not clear-cut, and individuals can show a preference toward either showing an N400 or P600 response to the same stimuli (Lee Osterhout, 1997). Even when presented with a violation of grammatical structure in the form of subject-verb disagreement within a sentence, a condition which is typically expected to result in a P600 response to reflect attempts at syntactic reassessment and repair, a subset of individuals show an N400 response (Tanner, 2013). Interestingly, a small proportion of individuals even showed both an N400 and P600 in response to violations, with variation in their

amplitude. Largely, however, there was a negative correlation between the amplitudes of the two responses, suggesting that participants preferentially showed one or the other (2013). These results suggested a spectrum of processing, which Tanner et al. (2013) posit suggest preferential attention to either semantic associative information (resulting in an N400), following a rule-based framework, or grammatical cues (resulting in a P600).

It has been suggested that an N400 might reflect rule-based, semantic processing in lower-proficiency individuals, which instead becomes a P600 to reflect a deeper understanding of grammatical structure in higher-proficiency individuals (Pakulak & Neville, 2010). However, subsequent findings complicate this interpretation (Tanner, 2013; Tanner et al., 2014). First, the preferential demonstration of either the N400 or P600 response has been seen both in native English speakers, and in individuals who learned English as a second language (Tanner, 2013; Tanner et al., 2014). Second, the preference toward showing one response or the other was not related to proficiency (Tanner et al., 2014). Instead, proficiency was related to the overall amplitude of the response shown, regardless of which component was evident. These findings illustrate a complex relationship between language proficiency and the role of the N400 and P600 in violation processing, and even language violations which are classically understood to be grammatical in nature can be processed in a semantic associative form in some individuals.

Despite that proficiency alone may not predict whether participants will exhibit an N400 or P600 response to specific types of language violations, it has been associated with characteristics of these components such as their latency and amplitude. For example, proficiency has also been related to processing of the grammatical content of a sentence, as the P600 shows a broader distribution, earlier onset, and larger peak amplitude for higher-proficiency individuals (Pakulak & Neville, 2010). Moreover, individuals with overall higher measures of proficiency have shown both an earlier onset, and reduced peak amplitude, in the

N400 response to violations of semantic relationships (Moreno & Kutas, 2005; Weber-Fox et al., 2003). Notably, these findings are in contradiction to those of Newman et al. (2012). At this time, the reason for this discrepancy is unclear. However, such inconsistencies highlight the need for further research in this area.

It has been suggested that differences in N400 amplitude reflects the degree to which individuals rely on the context of a sentence to predict its final word (those categorized as low-proficiency, resulting in a larger-amplitude N400), or instead focus on grammatical structure (Moreno & Kutas, 2005). However, both strategies may be used by individuals of high proficiency, and so a one-dimensional scale of proficiency may not be an appropriate interpretation of these findings. Instead, these findings present further evidence that individual variability in the use of divergent processing streams should be considered more deeply, and that both might be used simultaneously.

In addition to language proficiency, working memory appears similarly connected to processing strategy; Indeed, it has long been established that working memory capacity varies between individuals and may have implications for processing of linguistic content (Just & Carpenter, 1992). Specifically, increased working memory capacity has been related to elicitation of a P600 response, instead of an N400, when presented with grammatical violations (Nakano et al., 2010). This change is thought to contrast semantic reassessment of individual words (lower capacity) to higher-level phrase structure repair (higher capacity; Nakano et al., 2010). Similarly, Vos et al. (Vos, Gunter, Kolk, & Mulder, 2001) report that when presented with morphosyntactic violations, individuals with lower working memory span show stronger early anterior negativity and delayed central-parietal positivity, when compared with high-span individuals. Additional evidence suggests that individuals with faster working memory updating (measured using an *n*-back task) showed stronger central-parietal positivity (interpreted as a larger-amplitude P600)

when assessing syntactic and semantic information in sentences (Li, Peng, Liu, Booth, & Ding, 2014). fMRI evidence further shows differential recruitment of neuroanatomical structures for sentence processing depending on the working memory capacity of individuals (S. D. Newman, Malaia, Seo, & Cheng, 2013). This evidence demonstrates a link between aspects of language proficiency and cognition with presentation of ERP components that are commonly used to study language processing. Furthermore, they highlight the need to additionally consider cognitive attributes such as working memory that lie outside the obvious language processing domains when examining the effects of individual differences in cognition on language processing. Despite this, language proficiency and other aspects of cognition, are rarely accounted for in investigations of the N400 and P600. As such, a deeper investigation of the relationships between proficiency and cognition on these components will be necessary to understand their roles.

## 1.5. Modeling Individual Differences in ERP Research

Research investigating individual differences and their impact on ERP characteristics has begun to ask increasingly complex questions. For example, while at one time it was considered sufficient to investigate ERP responses to morphosyntactic violations across participants through a grand-averaged waveform. Contrasting conditions in this manner is a suitable use for the Analysis of Variance technique (ANOVA), which has been used to gauge the significance of differences between the mean amplitude of responses to grammatical and ungrammatical sentences (Moreno & Kutas, 2005; Pakulak & Neville, 2010). When evaluating a group of participants as a whole, or even a small number of subgroups, this approach can be used to characterize responses to each condition. However, it has since become apparent that not all participants show the biphasic N400-P600 response that has traditionally been expected (Tanner

et al., 2014). Instead, as discussed above, this biphasic response likely results from averaging of participants who preferentially show one response or the other together, resulting in an averaged waveform which shows the characteristics of both groups. Indeed, the majority of individuals show one response or the other (Tanner, 2013; Tanner & Van Hell, 2014). However, while the reason for this difference between individuals is not yet clear, it appears to be rooted in differences in cognition between individuals, as they selectively process linguistic information using either semantic associations or high-level grammatical rules. Other studies have categorized individuals as either low- or high-proficiency to investigate differences in responses between these two groups (Weber-Fox et al., 2003). However, whether participants are split into groups to test for the significance of differences separately, or whether a two-way ANOVA is used to account for differences on two dimensions (proficiency bins and sentence condition simultaneously), this approach precludes any description of graded changes between the two groups. Instead, distinct groupings must be made.

A more fine-grained characterization of the impact of individual differences in cognition or proficiency on ERP responses requires multivariate regression. This approach allows the user to investigate multiple individual differences simultaneously (e.g., several different axes of proficiency, and other cognitive factors, on continuous scales if desired). Indeed, multivariate regression has successfully been used to describe the impact of proficiency, familial handedness, and other factors both on the type of response a participant shows to grammatical violations, as well as the amplitude of responses (Tanner et al., 2014; Tanner & Van Hell, 2014). Not only does this approach allow the user to characterize multiple dimensions in terms of their effect on responses simultaneously, but it also allows for the use of continuous scales, and can describe interactions between dimensions (e.g., characteristics that are specific to, for example, left-handed males of low proficiency). This technique can be further refined through the use of

mixed models, which allow the user to specify both *fixed effects* (the types of factors described above, which are assumed to have a consistent effect) as well as *random effects.* The latter can improve the sensitivity of models by accounting for random variability in responses, either between individuals, or resulting from other factors that contribute to the response unpredictably (Barr, Levy, Scheepers, & Tily, 2013). To our knowledge, however, at present this technique has only been used in one study (Newman et al., 2012).

## 1.6. Limitations in Current Modeling Practices

The above studies have made important steps toward creating a framework which can be used to identify individual differences that impact the N400 and P600, as well as characterize this impact, in terms of latency, scalp topography, or the degree to which effects are seen at specific proficiency levels. However, several potential limitations of the techniques that have been used may be constraining the questions which can be asked. This dissertation will attempt to address each of these limitations using statistical modeling techniques which operate in fundamentally different ways to either the ANOVA or multivariate regression techniques that have been used. As a preface to a discussion of these limitations, it is critical to note that their existence may not necessitate a change in approach. That is, while the approaches previously used to investigate individual differences face constraints, these constraints may not limit interpretation of results in a meaningful way. However, developments in modeling approaches have resulted in a means for improving sensitivity to effects and complexity of interactions, among other added functionality, and so the potential for improvement of practices must be investigated. We will therefore be replicating and expanding on previous findings using a series of approaches, which are each intended to supplement the capabilities of current investigations.

The move toward multivariate regression models in recent years represents a step that will be required in subsequent research to properly account for the impact of all individual differences of interest, as well as how they interact to affect cortical responses (Tanner et al., 2014; Tanner & Van Hell, 2014). This is especially true given that, increasingly, individual metrics of grammatical ability vocabulary and other aspects of proficiency are incorporated into models, rather than one-dimensional metrics of overall proficiency (Newman et al., 2012; Pakulak & Neville, 2010). Given the benefits this framework provides, we see no clear advantages to relying on simpler approaches such as ANOVA. However, the move to mixed modeling techniques that incorporate both fixed and random effects further addresses the concern of random participant-specific variability, which is otherwise deemed to be noise when using fixed-effect models. Failure to properly designate this type of variability can make it difficult to identify subtle effects, and even result in errors in estimating the significance and size of effects which are found (Barr et al., 2013). This does not necessarily limit the types of research questions which can be asked, but instead the effects which can be uncovered in ERP data, especially where effects might be small. Despite this, the advantages of incorporating random effects into a regression model of individual differences' effects on ERP data have not yet been quantified. Therefore, we propose to investigate both approaches (fixed effect models and mixed models) to characterize the impact of including random effects of varying types on metrics of model quality. The manner in which the analytical techniques incorporate random effects will be discussed below.

We have discussed the need for a move from a single overall proficiency metric to a set of measures which account for the varied and dynamic nature of language-processing faculties. Indeed, this has been the approach taken in recent research (e.g., Newman et al., 2012; Pakulak & Neville, 2010). However, the way in which modeling techniques establish relationships between these measures and ERPs must be considered as well. In particular, moving from

analysis using ANOVA to linear regression has provided the means to describe a continuous, linear impact of individual differences on ERP responses. The result can be improved model fit, resulting in a reduced error term, and more reliable estimation of results under ideal circumstances. However, there is still a limitation in that this relationship must be linear, and to the extent that it is not, the model is inaccurate. That is, a linear model describes that any per-point proficiency score increase is associated with a specific voltage change in response amplitude, or millisecond change in timing, regardless of whether the largest changes in response may occur in specific portions of the proficiency spectrum. At present, it is unknown whether this is the case, and so whether this is a meaningful limitation and/or introduces error cannot be determined. However, given that the means to model nonlinear relationships exists, we propose to investigate the capabilities of two nonlinear modeling techniques in ERP data.

With the transition to the use of models which can include numerous aspects of language proficiency or any other measures of interest, the concern of multicollinearity arises. Multicollinearity occurs when the two or more predictor variables in a multivariate regression model are correlated in their ability to predict a response (in this case, multiple similar measures of proficiency attempting to predict ERP amplitude) (Bollinger, Belsley, Kuh, & Welsch, 1981). While this does not harm the fit of the model, it can introduce erratic changes in a predicted response given minimal changes in a predictor (Bollinger et al., 1981). The first of two solutions is to unify the related measures into a single predictor, either through averaging or some other means. However, as this defeats the purpose of looking at specific aspects of language proficiency *instead* of a general score, the only remaining option when using regression models is to include only those predictors which are most interesting and/or beneficial to the model. This process of model selection is not straightforward and many considerations must be made to arrive at the ideal set of predictors, which can be subjective given the specific research question.

However, future studies of individual differences will require some framework for addressing this problem, or otherwise a means to overcome it. The following chapters will attempt to do both through fundamentally different statistical approaches.

A related issue of data quality and distribution pertains to heteroscedasticity, which negatively affects all regression models as well as ANOVA (Davidson & MacKinnon, 1993). Briefly, heteroscedasticity refers to an inconsistency in variability either within groups of a categorical variable (e.g., there is more variability in the relationship between ERP response amplitude and low- than high-proficiency individuals), or along the scale of a continuous variable. Heteroscedasticity can be difficult to overcome, as options include either removal of problematic participants, portions of a predictor variable spectrum, or predictor variables altogether. However, some techniques are more susceptible than others to problems associated with heteroscedasticity, which will be explored in each of the techniques used in this dissertation.

The last issue to be discussed surrounding current statistical practices in the area of individual differences in ERPs surrounds the scalp topography of responses. Frequently, electrodes are considered on an individual basis in order to assess the significance of a difference between response amplitudes in two conditions. This is commonly done to evaluate the differences between responses to words that either follow or disrupt a sentence's grammatical structure. This has been done to clearly demonstrate an N400 response to semantic violations at individual electrodes (Moreno & Kutas, 2005). In addition, the topography of a response can also be visualized using response amplitude across numerous electrodes for any given time point, a series of time points, or averaged over a time window. This process has been used to identify characteristic differences of the P600 topography in low- and high-proficiency individuals (Pakulak & Neville, 2010). This has the advantage of presenting a full view of the

scalp, but with less information regarding the latency and time course of a response. Another approach is to average electrodes in a close proximity together into a region of interest (ROI), with the assumption that any characteristics of the response that are similar across electrodes will remain through the averaging process, and those which are not will be reduced (Newman et al., 2012). Establishing regions of interest can produce a more consistent waveform, and aid interpretability. However, divisions must be made without specific knowledge of the response's topography, and thus are not likely to perfectly capture the effect. Ideally, divisions could be made which are data-driven and representative of the full scalp, benefiting from the advantages of each of these methods.

The above points outline a variety of constraints that currently-used statistical methods place on framing of research questions and interpretation of findings. It is possible that some of them cannot be overcome using available techniques, but the capacity to do so will be investigated from several different analytical approaches. The following sections will outline the three techniques that will be used in this dissertation, in order to discuss how they will address the known issues, specifically in the context of modeling ERP data.

## 1.7. Linear Mixed Modeling

The first modeling technique we used was intended to closely follow the methods in related research discussed above. First, we attempted to replicate findings which relied on multivariate regression (Tanner et al., 2014; Tanner & Van Hell, 2014), with the exception that we implemented random effects similarly to Newman et al. (2012) to account for unpredictable variability from two sources: Random variability in response amplitude between participants, and in the scalp distribution of the response between participants. For this analysis, we used linear mixed effects (LME) modeling. As discussed above, this technique models the relationship

between fixed effects and a response, as traditional multivariate regression models, but can also estimate the influence of random effects (Jiang, 2007). As a rule of thumb, Green and Tukey (1960) suggest that effects should be modeled as fixed if the sample exhausts a known population, but random if it includes only a subset (i.e., individual participants). Perhaps less specifically, LaMotte and Roy (LaMotte, 2014) describe a random effect as any that is assumed to be a realized value of a random variable. Mechanically speaking, however, random effects are simply estimated using a different process – linear unbiased prediction, rather than maximum likelihood – which allows for differences in variability across a predictor variable (Robinson, 1991). In terms relating individual differences to ERP responses, we must use this definition to assume that differences in language proficiency or other measures of cognition should be modeled as fixed effects, given that the degree of variability should be relatively constrained across participants. Indeed, this has been the case in all prior research in this area.

Conversely, it could be argued that individuals should likely vary in the degree of variability in response amplitude for any ERP study. This variability may arise for numerous reasons, including – but not limited to – differences in scalp conductivity, biochemistry, neuroanatomy, or any other physiological traits unique to an individual which cannot be modeled. Indeed, this assumption has been made in the past, and modeling random variability at the individual level is an accepted means of improving model quality (Barr et al., 2013). Just as random variability might be expected at the individual level, one might expect a degree of random variability in the scalp distribution of responses between individuals. This should not be confused with the assumption that variability in scalp distribution is inherently random, which would invalidate findings of proficiency effects on distribution. Instead, a random effect of scalp distribution across individuals would allow for variance in distribution between trials that is specific to participants. A standard instantiation of the general linear model (GLM), such as the

multivariate regression approaches used by Tanner et al. (2014; 2014) is well-suited to investigating fixed effects, but is not capable of estimating random variability from either of these two possible sources. The result is that all participants are predicted to have an equivalent variability (both overall and in terms of scalp distribution), and any variability that is observed around that assumption is considered noise, which is added to the error term. Moreover, heteroscedasticity may result. Both of these outcomes may render a model insensitive to effects and/or invalidate its findings (Jiang, 2007).

Regression models calculate the significance of fixed effects in relation to the magnitude of the error term, and so categorizing unpredictable variability as error occurs at the detriment of sensitivity to potentially subtle effects. That is, when the magnitude of explainable variability (e.g., proficiency-related effects) is outweighed by that of unexplainable variability (e.g., random variance in response amplitude due to physiology which cannot be modeled as a fixed effect), model sensitivity decreases, and underlying effects may not be detected. While subsequent chapters will address this concern in more detail, it is important to realize that adding random effects to a model can help to attenuate unpredictable variability that arises from known sources. The primary benefit of adding random effects through LME then is improved model accuracy and sensitivity, as well as reduced error rates (both false positive and false negative).

In addition to modeling random effects, LME was used to determine the ideal model selection process in a way that may generalize to other ERP studies of individual differences. As described above, this primarily consisted of determining which individual differences should be included as predictor variables in our regression model, both in terms of their significance to the research and the relatedness of the measures. Specifying the ideal model is not specific to LME, and investigations that only use fixed effects can still benefit from determining which language proficiency or other cognition measures are appropriate for inclusion in a model. Moreover, the

metrics which are commonly used to evaluate model quality and interrelatedness of variables are applicable to all linear modeling techniques, and so these findings may result in a framework that can benefit ERP studies more generally. These metrics will be discussed in greater detail in following chapters. The process of model selection was explored with the aim to reduce multicollinearity among predictors, further reducing error rates and improving model accuracy, while focusing on predictors that should yield the clearest results. Determining the ideal set of predictors (and the complexity of their interactions) also provided a foundation for subsequent analyses, as each approached incremental improvements on the same base model in different ways.

## 1.8. Generalized Additive Mixed Modeling

The sensitivity of a model to an effect primarily depends on the ability of a model to fit to the association between two variables. To the extent that this association is not a linear one, nonlinear modeling techniques may produce a more accurate fit, and a model which more ideally generalizes to a population. As the name suggests, all extensions of the general linear model are limited in their capacity to conform to nonlinear effects. For example, if differences in the amplitude of an N400 response exist primarily between individuals at the lower end of a proficiency spectrum (that is, higher-scoring individuals produce more homogeneous responses), then we can determine that the relationship between proficiency and N400 amplitude is nonlinear. However, fitting this relationship with a linear best-fit slope both fails to capture the nuance of this dependence, and over- or underestimates the predicted N400 amplitude to some degree at nearly any proficiency level. This concept is illustrated in **Figure 1.1**, which shows a simulated set of observations fit to two such functions. The result is increased error (residual

variance), and a linear function which only accurately estimates observations at mid-range scores.

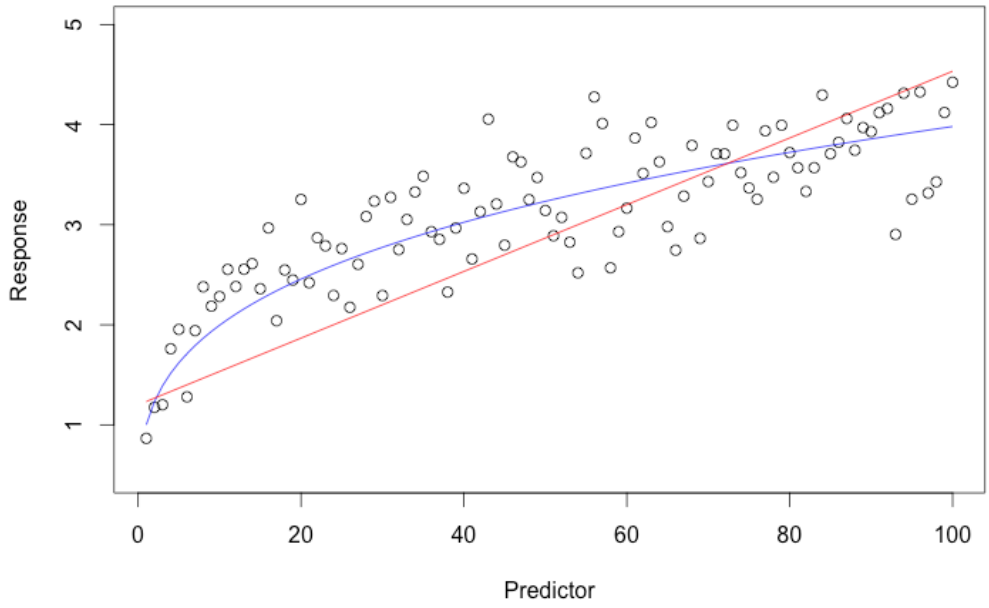


**Figure 1.1** Fit of a linear function (red) and a nonlinear function (blue) to simulated data which follow a nonlinear curve $Y = X^{0.3}$. While both achieve an overall fit to the data points, the linear function produces estimates that are consistently lower than observations at lower predictor values, while exceeding observations at higher predictor values.

A combination of related problems arise from the scenario depicted in **Figure 1.1**. For example, as deviance from the predicted slope is necessarily considered error variance, this results in both increased error variance (again at the detriment of significance of effects), and non-normally distributed residuals (i.e., heteroscedasticity). Again, while we will examine each of these concerns in more detail in subsequent chapters, each represents unexplored bias introduced into findings which can potentially reduce sensitivity to effects or invalidate a model's findings altogether, depending on the degree to which a nonlinear dependence truly exists.

In order to evaluate whether the relationship between ERP response amplitude and language proficiency and other cognitive measures is better-modeled using nonlinear functions, we used Generalized Additive Mixed (GAM) modeling. This is a technique which linearly combines estimates for separate predictor variables, similarly to LME. However, functions need not be linear – Instead, smooth polynomial functions are fit to observations to describe nonlinear associations (Hastie & Tibshirani, 1990). While maintaining the benefit of random effect specification described when using LME, this approach uses an entirely different approach to model fit, which will be explored in greater detail in subsequent chapters. Moreover, the computational differences between LME and GAMM have implications for which model quality metrics are appropriate to use, and how heteroscedasticity is handled (Simon N. Wood, Pya, & Säfken, 2016). Both of these topics will be discussed in greater detail in the chapters pertaining to these two techniques.

Using this technique we examined first whether models are improved through specifying nonlinear effects/interactions, and second the ideal complexity of this relationship (i.e., the degree of nonlinearity allowed). This second goal was to strike a balance between describing the overall trend in as much detail as possible while avoiding fitting to characteristics of the data set which may be specific to the present sample (i.e., over fitting). There is no gold standard in this process, as it cannot be known beforehand whether variance is (or should be) generalizable to a population. Together, these analyses aimed to answer not only the question of whether nonlinear relationships exist between language proficiency and response amplitude, but also whether it is important or viable to model them as such.

## 1.9. Conditional Inference Random Forest Modeling

The final modeling technique which will be used in this dissertation is conditional inference random forest modeling (CForest). LME and GAMM are both approaches which arose from the GLM, and as such they both linearly combine estimates for predictors, and interactions between predictors, with an error term to fit to a set of observations. It is for this reason that both are considered additive modeling solutions. Conversely, CForest begins with the observed data and through a series of permutation tests identifies where predictors are associated with systematic variations (Strasser & Weber, 1999; Strobl, Malley, & Tutz, 2009). Therefore, predictors are not assigned estimates, and there is no error term. Instead, data are recursively subdivided into increasingly small subsets, provided that doing so will result in a significant difference between the means of the response variable in each, and that this difference is associated with a predictor. This approach makes far fewer assumptions of a data set. For example, as there is no error (residual) term, normality of residuals is not a concern (Strobl et al., 2009). Computation of CForest models will be explained in greater detail in subsequent chapters.

One notable strength of CForest is that it does not require *a priori* definition of regions of interest, but can instead deduce which electrodes show an effect – and which do not – through recursive partitioning. Ideally, this should result in data-driven groupings of electrodes which more closely capture the distribution of a response. Conversely, each of LME and GAMM relies on *a priori* specification of the region in which the presence of an effect will be evaluated. This is commonly done through division of the scalp into groupings of electrodes, or regions of interest (ROIs) that are evaluated as a whole. This can be done either through averaging the response over the electrodes in an ROI, or through including ROI (but not electrode) as a term in

a model. Grouping electrodes reduces electrode-specific signals, but it cannot be known whether *a priori* groupings will ideally capture a component's distribution. For example, the peak amplitude of a component may be at the division of two ROIs, meaning that it will appear somewhat mitigated in each averaged group of electrodes. This may also result in failure to detect an effect altogether, or mischaracterizing a response's topography to appear more widespread than it truly is. Ideally, this effect could be better-captured with an ROI that centers on its peak amplitude, and exceeds more specifically to the boundaries of the response's spatial distribution.

As an alternative to pre-selected ROIs, clustering algorithms can be used to group electrodes that show similar ERP waveforms (Pernet, Latinus, Nichols, & Rousselet, 2015a). This approach has proven capable of grouping ERP responses in a data-driven way that can reduce user bias. One limitation, however, is that clustering may not reflect differences in scalp topography between subgroups of respondents (i.e., low- and high-proficiency individuals) unless those groups are identified beforehand. Therefore, when the influence of some predictor is on the spatial distribution of a response, this type of clustering approach may not be ideal. In addition, GAMM is able to circumvent this limitation by modeling the topography as a smooth term (or set of terms) over a spherical model of the head, using electrode coordinates as a predictor. However, the present analysis used the electrode groupings that were used in our LME analysis and in previous research to aid comparison of findings (A. J. Newman et al., 2012).

In practice, the effects of interest in the present study (the N400 and P600) frequently have a central-parietal distribution which is captured relatively well using an equal grouping of electrodes into regions, as has been done by Newman et al. (2012). However, differences in the extent of an effect can be difficult to evaluate. Moreover, square groupings of electrodes obviously fail to perfectly encapsulate components which are circular in their topography, where

the square is either so large as to include corners which do not show the response, or so small as to ignore its edges. We propose that CForest can be used to define ROIs by deducing the distributions of responses. The data-driven nature of CForest's hypothesis testing framework was leveraged to derive ROIs which conform to the circular shape of the N400 and P600, and reach a distribution of electrodes that appropriately describe its topography (Strasser & Weber, 1999).

In addition to these advantages, CForest has proven capable of detecting nonlinear relationships in neuroimaging data previously (McWhinney, Tremblay, Chevalier, Lim, & Newman, 2016). Indeed, the recursive partitioning framework does not require any assumption of linearity between predictor variables. While this approach has numerous advantages, limitations and considerations to its use which will be described in later chapters, it represents a step forward in data-driven descriptive analytics.

## 1.10. Preface to Investigations

The investigations in the following chapters each take a different approach to analysis of ERP data acquired during a sentence judgment task, which was designed to elicit cortical responses to violations of either expected semantic associations, or morphosyntactic phrase structure. Sentences were either well-formed or contained a single type of violation. Semantic violations replaced the final word of a sentence with one that is semantically implausible in the context of the sentence. Conversely, sentences containing grammatical violations contained nouns that followed expectations of semantic associations, but disrupted the grammatical structure of a sentence. Participants were each evaluated in several areas of language proficiency and working memory capacity to determine whether these were related to differences in the characteristics of the elicited components that were elicited by the two types

of violations. This task and all evaluations will be described in greater detail in the following chapter, which outlines procedural details, data acquisition and pre-processing.

Chapters addressing the limitations and concerns that have been outlined above will be presented, each focusing on separate statistical techniques. These chapters aim to expand on the data set through different means of analysis, each with strengths in unique areas. Each chapter will consider the appropriate use cases and potential limitations of the approaches being discussed, assessing and possible improvements in how these techniques can aid with the interpretation of ERP data analysis. Lastly, we will look critically at any benefits that these techniques can provide to researchers, and most importantly any evidence that they should be considered for future use in ERP research.

The goal of these investigations was to first apply best-practices to develop a framework in which relationships between individual differences and ERP characteristics could be assessed, and second to evaluate the ability of alternative modeling techniques to overcome some of the limitations associated with currently-used approaches. While our aims were methodological in focus, the nature of the data used led to specific predictions about the outcomes. First, in line with previous research, sentences that contained semantic violations, when contrasted with well-formed sentences, were hypothesized to elicit an N400 response at the onset of the semantically incongruent word. This response was predicted to show the highest amplitude in participants with lower measures of proficiency, given that N400 amplitude is thought to index the mental effort required to draw semantic associations (Marta Kutas & Federmeier, 2011; Moreno & Kutas, 2005; Weber-Fox et al., 2003). In particular, participants' vocabulary scores were hypothesized to be most strongly related to N400 response amplitude (Newman et al., 2012).

Second, it was hypothesized that sentences that contained phrase structure violations, when contrasted with well-formed sentences, would elicit a P600 response. This response was predicted to be strongest in participants with higher scores in proficiency-related ID measures, as has been demonstrated in previous studies (Pakulak & Neville, 2010). In particular, measures of grammatical ability were expected to be most strongly associated with P600 amplitude (Pakulak & Neville, 2010).

Concerning the topography of responses, to the extent that ID measures influenced the amplitude of responses, it was expected that their influence would be strongest where the peak of responses was maximal (e.g., over the central parietal scalp region). Conversely, to the extent that differences in ID measures were associated with differences in the spatial extent of a response rather than amplitude, the influence of ID measures was expected to be strongest at the boundaries of the response topography. Beyond these predictions, hypotheses that were specifically related to any of the three analysis techniques will be addressed in the chapters devoted to those methods.

# Chapter 2: Data Collection and Pre-Processing

As a number of details regarding the collecting, cleaning and preparation of data pertain to each of the analyses described in the previous chapter, all such details are outlined here. These include a description of the participants included in the analyses, the questionnaires and assessments used to evaluate language proficiency and working memory capacity, and a description of the sentence judgment task that participants completed. In addition, basic pre-processing procedures that applied to all later statistical analyses are described here.

## 2.1. Participants

The experiment collected data from 40 native English-speaking individuals, of which 33 were included in analyses (8 male, mean age = 28.72 ± 9.49 years, range = 20 - 57). Two participants were excluded due to excessive noise in EEG data, and five due to errors in data acquisition. Three of the seven excluded participants were male. All participants were right-handed, as indicated by the Edinburgh Handedness Inventory (Oldfield, 1971). Participants had normal or corrected-to-normal vision, and no self-reported history of neurological disorders. This research was approved by the Dalhousie Health Sciences Research Ethics Board. All Participants provided informed consent ethics board guidelines. Each participant was paid $30 upon completion of their participation.

## 2.2. Questionnaires

Due to the range of ID measures investigated, participants were required to complete a number of questionnaires and tasks. An outline of all assessments is provided in **Table 2.1**, with a brief description of the domain or metric addressed in each. The assessments are additionally described in greater detail below. Notably, several measures were not included in our modeling

procedure. The Edinburgh Handedness Inventory was used, but not included, to ensure that all participants were right-handed. Only minimal variation in handedness was therefore expected and was not predicted to significantly influence our model. The Language Experience and Proficiency questionnaire was used to evaluate first and second language background, but more specific measures of proficiency (e.g., TOAL-3) were expected to be more related to variability in cortical responses, particularly on the dimensions the task was designed to evaluate. Lastly, the Index of Learning Styles questionnaire and the Flanker-Simon executive attention task were available but were peripheral to our literature review and so no specific hypotheses could be drawn regarding these assessments.

***Table 2.1*** *Questionnaires and tasks completed by participants. Those in bold were included in models.*

| Questionnaire or task name | Assessment content |
|---|---|
| **Edinburgh Handedness Inventory** | Left vs. right handedness |
| **Language Experience and Proficiency Questionnaire (LEAP-Q)** | First and second language background |
| **Index of Learning Styles Questionnaire** | Learning preferences on four dimensions |
| **Test of Adolescent Language-3 (TOAL-3)** | Grammar comprehension and vocabulary |
| **Test of Word Reading Efficiency (TOWRE-2)** | Test of word reading efficiency |
| **AzBio Sentence Task** | Test of speech comprehension |
| **Operation Span (OSpan)** | Working memory capacity (visual) |
| **Listening Span (LSpan)** | Working memory capacity (auditory) |
| **Flanker-Simon Task** | Executive attention |
| **Sentence violation identification** | English semantic and phrase structure |

Participants completed a set of questionnaires using the online LimeSurvey software (LimeSurvey Development Team, 2012), which included the Edinburg Handedness Inventory (Oldfield, 1971), the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007), the Index of Learning Styles (Felder & Soloman, n.d.) and the Adult Reading History

Questionnaire (Lefly & Pennington, 2000). Questions were primarily multiple choice, with some short answer, and were intended to evaluate a wide range of individual differences in handedness, first and second language background, learning style preferences, and demographic information.

## 2.3. Language Assessments

Two facets of native language proficiency were evaluated using three assessments. First was the Test of Adolescent and Adult Language-3 (TOAL-3; Hammill et al., 1994), of which we used three subtests: the Listening/Vocabulary subtest had participants match an auditory stimulus (word) to one of four pictures to measure receptive semantic abilities; the Listening/Grammar subtest had participants determine which of three spoken sentences presented together has a different meaning, to assess receptive morphology and syntactic abilities; the Speaking/Grammar subtest had participants listen to and repeat increasingly complex sentences to assess expressive morphology and syntactic abilities. The scores on each subtest were retained as individual predictor variables, but were scaled from 0-100 rather than their native scales to aid interpretability. This scale was used to keep consistency among scores associated with high- and low-scoring participants across measures, and to mitigate convergence errors that were encountered when using LME which were caused by differing predictor scales. Moreover, while mean-centering predictors has been advocated as a means of reducing multicollinearity in regression models, it has been shown to have no impact on model term estimates or significance (Dalal & Zickar, 2012; Kromrey & Foster-Johnson, 1998). Moreover, this transformation has been shown only to render multicollinearity undetectable (Belsley, 1984). Therefore, no such transformation was applied to our ID measures scores.

The second proficiency assessment was the Test of Word Reading Efficiency, 2nd Edition (TOWRE-2; Torgesen et al., 1999). This assessment used two separate tests to measure the number of real words or non-words in a list that participants can read out loud within 45 seconds, yielding, respectively, sight word efficiency and phonemic decoding efficiency scores. A composite measure of the two known as the Total Word Reading Efficiency was used as an indicator of overall reading efficiency. The last proficiency assessment was the AzBio Sentence List (Spahr et al., 2014), which evaluated speech perception when masked by noise. Participants were asked to repeat any complete sentences that were heard amongst 'cocktail-party' (multi-talker babble) environmental noise. Participants' verbal responses to both the TOWRE and AzBio tasks were audio recorded for later scoring.

## 2.4. Cognitive Assessments

Participants completed several other computer-based cognitive tests using custom-written software. The Operation Span task is a measure of working memory capacity (OSpan; Unsworth et al., 2005). Participants were presented with basic math equations of three single-digit terms (e.g., 2 × 6 = 3), and were then asked to judge whether a provided answer was correct or incorrect. Each judgement was followed by a single letter, and over the course of several trials participants were instructed to remember these letters in order. Following a number of equations (one to seven, with one letter for each equation), participants were asked to repeat the letter sequence from memory. Sequences of equation/letter pairs began short (3) and continued until the maximum length was reached (7).

A second measure of working memory capacity was a listening span task (LSpan; Daneman and Carpenter, 1980). In this task participants were presented with a series of auditory sentences. After each sentence, a comprehension question was asked, and participants were

instructed to remember the final word of each sentence. Recall of the word list was evaluated verbally after the final sentence in a series. The structure of this task closely followed that of the OSpan task, with the length of the series of sentences (and the number of words to be remembered) gradually increasing until the maximum length was reached (3) or participants failed 7 consecutive trials. For both tasks, the length of the longest accurately-recalled sequences was used as working memory span.

The Flanker-Simon task was used to measure executive attention; it combines aspects of the Flanker (Eriksen & Eriksen, 1974) and Simon (Simon & Barbaum, 1990) executive functioning tasks. The Flanker-Simon task measures both independent executive function systems associated with each task as well as conflict adaptation. However, as described above, this task was peripheral to our investigation and was not included in our final analysis.

Participants received both verbal and written instructions on how to complete the task; they were informed to indicate the colour of the centre pinwheel as quickly as possible without sacrificing accuracy. They were instructed to direct their gaze toward the fixation point in the centre of the computer screen. After 1 s of fixation, stimuli were randomly presented. Participants used the trigger buttons to indicate the colour of the centre pinwheel. Participants were given 1 s to make a response, otherwise they received a warning message (“Too slow!”) and the next trial proceeded. If a response was made, participants were given visual feedback regarding their reaction time (in ms) and the colour of the text indicated the trigger they pressed. The Flanker-Simon task was composed of six blocks: One practice block and five experimental blocks. The task lasted approximately 15 minutes.

## 2.5. Sentence Task

The main outcome measure was ERP responses to a sentence reading task. A total of 640 sentences were developed, each of which fell into one of three categories. This categorization mirrored that used by Newman et al. (2007):

1. Morphosyntactic condition: Sentences containing past-tense violations (80) were contrasted with matched well-formed sentences (80). For example, *Last week she **asked** to watch the concert* / *Last week she **ask** to watch the concert* [1]*. Data from these sentences was not included in the analyses presented here.
2. Semantic condition: Sentences in which the final word was replaced with one that is semantically implausible (80) were contrasted with matched semantically plausible sentences (80). For example, *The farmer spends the morning milking his **cows*** / *The farmer spends the morning milking his **book*** *.
3. Phrase structure condition: Sentences that violated acceptable phrase structure (160) were contrasted with matched well-formed sentences (160). This condition contained twice the number of sentences when compared with the others to accommodate Steinhauer, White & Drury's (2009) suggested counterbalancing of the words preceding the violating word. The result was two control and two violation variants for each sentence. For example, *Bob likes to tell some stories at night* OR *Bob likes some stories to tell at night* / *Bob likes to stories some tell at night* * OR *Bob likes some tell to stories at night* *.

[1] Astrisks (*) indicate a sentence with a violation

Sentences from the above categories were randomly divided into groups as follows: 1) Sentences in the morphosyntactic category were divided into two groups, each containing 40 sentences with violations and their well-formed matches. 2) Sentences in the 'semantic' category were divided into two groups in the same manner. 3) Sentences in the 'phrase structure' category were divided into four groups in the same manner, rather than two groups, due to having twice the number of sentences to begin with. These subsets of sentences were combined into four sentence lists, where each list contained sentences from each category by using all four permutations of the groups from the first two categories (two 'morphosyntactic' groups by two 'semantic' groups), with one of the four 'phrase structure' groups appended to each. Each sentence list therefore contained 240 sentences.

## 2.6. Procedure

Participants attended two sessions, approximately 2.5 hours each. In the first, they provided informed consent, competed all questionnaires, the TOWRE-2, the TOAL-3, the LSpan, and OSpan tasks. During the second session, participants completed the AzBio and Flanker-Simon tasks, and then the sentence reading task while EEG data were recorded.

When completing the sentence task, each participant was only exposed to one of the four sentence lists described above. Participants were shown sentences one word at a time using rapid serial visual presentation; each word was visible for a random period of time selected from a uniform distribution, which ranged from 325 to 425 ms, with no delay between words. Stimuli were presented within a square of constant size, which was provided as a cue to participants, during which time they were instructed not to blink. At sentence completion, a blank screen was presented for 1000 ms, after which time participants were shown a 5-point Likert scale on the screen and asked to rate the "quality" of each sentence using a numeric

keypad (1 = very bad, 5 = very good). Beyond being instructed to rate sentence quality, participants were not given specific instructions about how to assign ratings. An additional blank screen was shown for 1000 ms before the next sentence was presented. Responses and reaction times were recorded for each trial, and breaks were provided every 15 sentences, which lasted until the participant pressed a button to continue. There were four additional experimenter-mediated breaks for any necessary adjustment of equipment, such as checking and lowering EEG electrode impedances.

## 2.7. EEG Acquisition and Pre-Processing

EEG data were collected using a 128 channel HydroCel Geodesic Sensor Net connected to a NetAmps 100 amplifier (Electrical Geodesics, Inc., Eugene, OR). Data were acquired using NetStation software (version 4.3, Electrical Geodesics Inc., Eugene, OR) at a 500 Hz sampling rate, using a 0.01-100 Hz online band pass filter, and referenced to electrode Cz.

All EEG electrode impedances were lowered below 100 kΩ prior to recording, which is an appropriate level according to the input impedance of the NetAmps amplifier. After data recording was completed, a 0.1-30 Hz band pass filter was applied in the NetStation software, and then data were exported from the NetStation to binary format and all further pre-processing was completed using EEGLAB (v12; Delorme and Makeig, 2004). Data were re-referenced to the average of the two mastoid electrodes. Epochs were generated beginning 200 ms prior to, and 1000 ms following, the onset of target words. Channels or individual trials showing excessive noise were removed, and independent component analysis (ICA; Jung et al., 2000) was used to remove well-defined ocular or muscular artifacts, on the basis of topography, power spectra, and distribution over time and across trials (Makeig, Bell, Jung, & Sejnowski, 1996). After ICA correction was applied, data for channels that had been removed prior to ICA

were interpolated using a spherical spline algorithm, and all channels were baseline-corrected by subtracting the mean amplitude of the 200 ms prior to stimulus onset from each electrode. As an alternative to baseline correction, the baseline amplitude for each electrode or ROI in each participant might be modeled as a random effect. However, the present analyses instead used baseline correction to account for inter-trial variability, which was not included as a model term.

ERP epochs were time-locked to either the onset of the violating word (violation sentences) or the word in its place in the well-formed matching sentences (control condition). For visualization of the ERP waveforms, violation effects were computed by subtracting control from violation sentence epochs.

## 2.8. Statistical Analyses

For statistical analysis, mean amplitude across a priori specified time windows was computed for each electrode on every trial, and then exported in text format for subsequent statistical analysis. Electrodes were grouped into nine regions of interest (ROIs) for statistical analysis; electrodes not falling within these ROIs were not analyzed. These groupings were done on a 3x3 grid, including an anterior-posterior axis (anterior, central, posterior) and a left-right axis (left, midline, right). The electrodes included in each ROI, and their positions, are depicted in **Figure 2.1**.

All further statistical analyses were specific to the three approaches outlined in the previous chapter. Therefore, detailed descriptions of these analyses will be provided when describing the methodology of each approach in subsequent chapters.

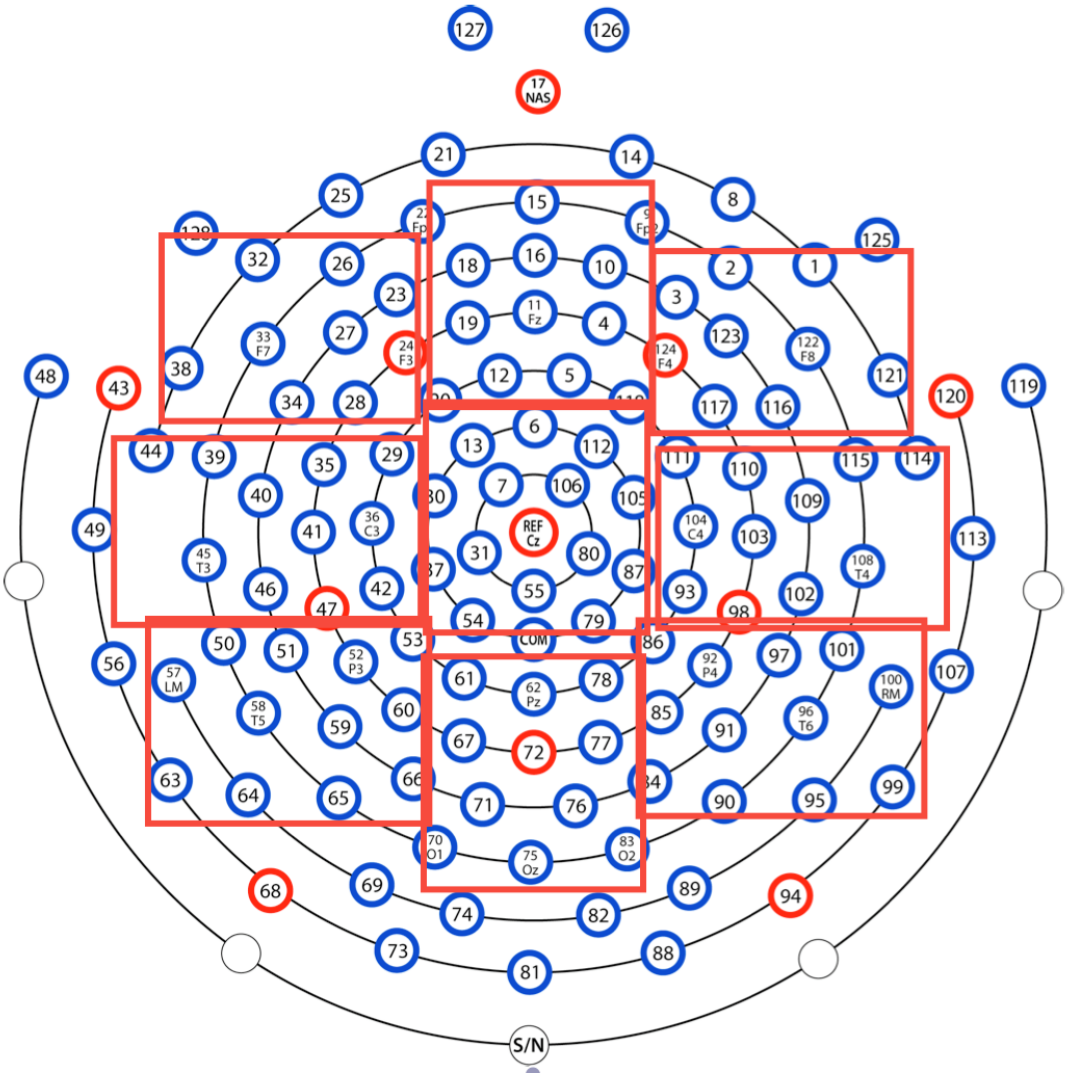


***Figure 2.1*** *Electrodes in each region of interest (ROI). ROIs are arranged on the* y *axis (anterior, central, posterior) and the* x *axis (left, midline, right) to produce each of nine (e.g., anterior midline).*

# Chapter 3: Characterizing the Data - Violation Effects and Individual Difference Measures

## 3.1. Violation Effect Time Course and Topography

General findings surrounding the effect of semantic and phrase structure violations which pertain to all of the statistical approaches used will be described briefly. In addition, we will outline the results of the individual difference measure assessments and overall relationships between these scores and response amplitude, independent of any specific modeling technique. The following chapters will each investigate these findings in more detail using the modeling techniques which have been described.

Given that the effects of interest for each sentence type are expected to be maximal in central-parietal scalp regions, violation effects are shown for channel 55, as seen immediately posterior to the reference channel in **Figure 2.1**. The violation effects for each sentence type are shown in **Figure 3.1**. The electrode used for demonstration of violation effects was selected due to being central in the topography of the response. Following a 200 ms baseline, semantic violations showed a notable (approximately -4 μV) negative deflection, which was maximal at between 300 and 500 ms, as predicted. This was immediately followed by a positive deflection of almost equivalent magnitude (3.5 μV), which was prominent from approximately 500 ms until the end of the epoch. Our later time window (600-800 ms) coincides with the peak of this deflection.

Responses to phrase structure violations did not include the early negative response seen to semantic violations, but showed similar positivity (approximately 3 μV in amplitude) peaking at between 500 and 900 ms following the onset of the violating word. Notably, peak

amplitude of this deflection occurred slightly later than in response to semantic violations. In addition, the peak of the waveform is somewhat weaker.

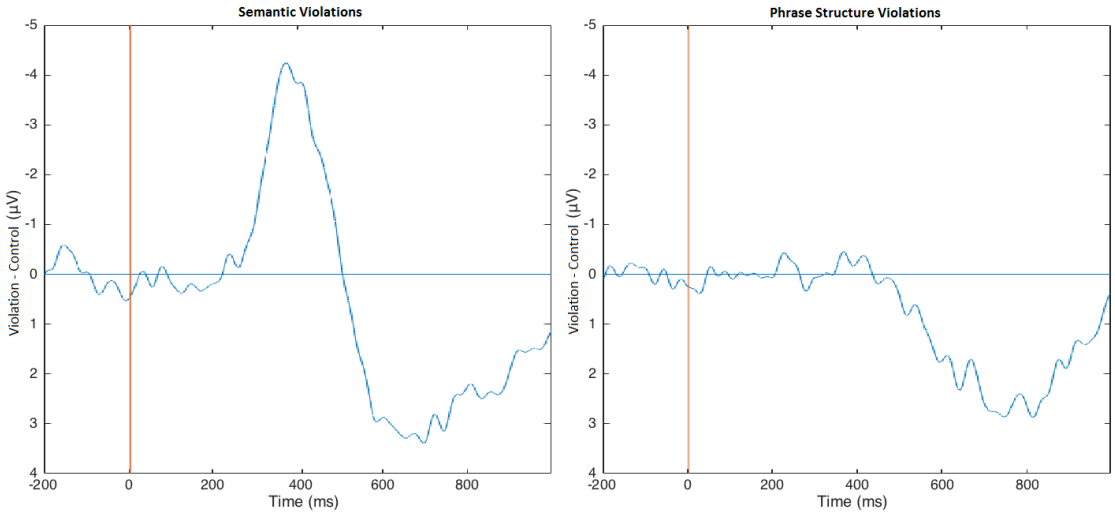


***Figure 3.1*** *Violation effects (averaged violation sentences – averaged control sentences) for each of semantic and phrase structure violations, at channel 55. Time 0 ms indicates onset of the word which violates the semantic or grammatical structure of the sentence.*

The scalp topographies of these violation effects are shown in **Figure 3.2**, averaged over each of the time windows of interest (300-500 and 600-800 ms) for each sentence type. Mirroring the time course of the electrode in **Figure 3.1**, for the semantic condition the earlier time window showed a negativity predominantly over central and posterior scalp regions, consistent with the predicted N400; notable positivity was seen in anterior electrodes. The subsequent (600-800 ms) response was positive, with a more central distribution. Phrase structure violations showed little violation-control difference in the earlier (300-500 ms) time window, in keeping with the time course shown in **Figure 3.1**. The later time window showed positivity with a similar scalp distribution as for semantic violations, but of somewhat weaker amplitude.

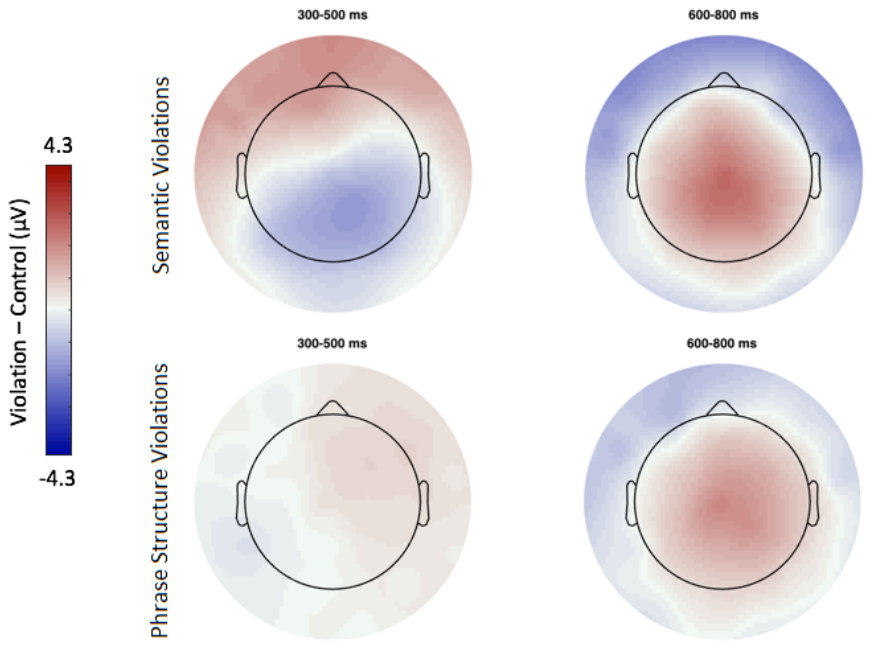


***Figure 3.2*** *Scalp topography of violation effects (averaged violation sentences – averaged control sentences) for each of semantic and phrase structure violations, averaged over the time windows of interest (300-500 ms and 600-800 ms).*

## 3.2. Individual Difference Measures

Prior to evaluating the relationships between each of the ID measures and the amplitude of responses to violations of either sentence type, we investigated the distribution of scores on each of the ID measures to ensure that there were no outliers and address any skew in their distribution. These results are shown in **Figure 3.3**. The scores of two participants on the AzBio task met our criteria for outliers (any scores more than 3 standard deviations above or below the mean), and so those scores were removed from all tables, figures and analyses below, as well as from all subsequent analyses in the following chapters.

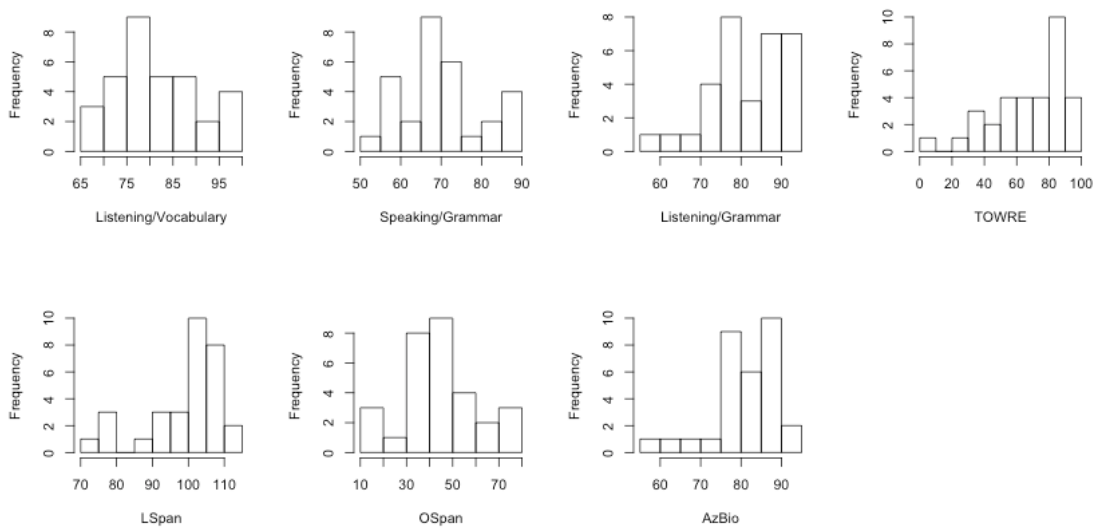


**Figure 3.3** Histograms of participants scores on each of the ID measures evaluated in the present study.

While the sample size (n=33) was unlikely to produce strongly normal distributions of ID measure scores, many of the measures still provided a reasonable separation of low- and high-scoring participants. Concerning the LSpan and particularly the AzBio tasks, participants were skewed toward higher score ranges more so than for other tasks, suggesting that finding differences in response amplitudes for lower- vs. higher-scoring participants on these measures may be less likely. Descriptions of the distributions of these ID measure scores are additionally outlined in **Table 3.1**.

***Table 3.1*** *Distribution of ID measure scores, including the number of unique values where multiple participants may have scores identically, the range of scores, the mean and standard deviation.*

| | UNIQUE VALUES | MINIMUM SCORE | MEAN SCORE | MAXIMUM SCORE | STD. DEVIATION |
|---|---|---|---|---|---|
| **LISTENING/VOCABULARY** | 12 | 66 | 82 | 100 | 11 |
| **SPEAKING/GRAMMAR** | 11 | 50 | 71 | 90 | 13 |
| **LISTENING/GRAMMAR** | 12 | 57 | 78 | 94 | 12 |
| **TOWRE-2** | 22 | 6 | 67 | 99 | 25 |
| **LSPAN** | 18 | 71 | 98 | 112 | 12 |
| **OSPAN** | 22 | 15 | 45 | 75 | 16 |
| **AZBIO** | 23 | 59 | 80 | 91 | 9 |

Participants' scores on each of the individual difference measures were compared with their epoched condition contrast (violation – control) for each of the two time windows (300-500 ms and 600-800 ms) in each of the two sentence types (semantic violations and phrase structure violations), averaged over trials and electrodes. This step was performed on pre-processed scalp voltage recordings, rather than any model estimate, to provide a frame of reference for each of the modeling techniques that was used in the following chapters. The results of these comparisons are shown for semantic violations in **Figure 3.4**, and for phrase structure violations in **Figure 3.5**.

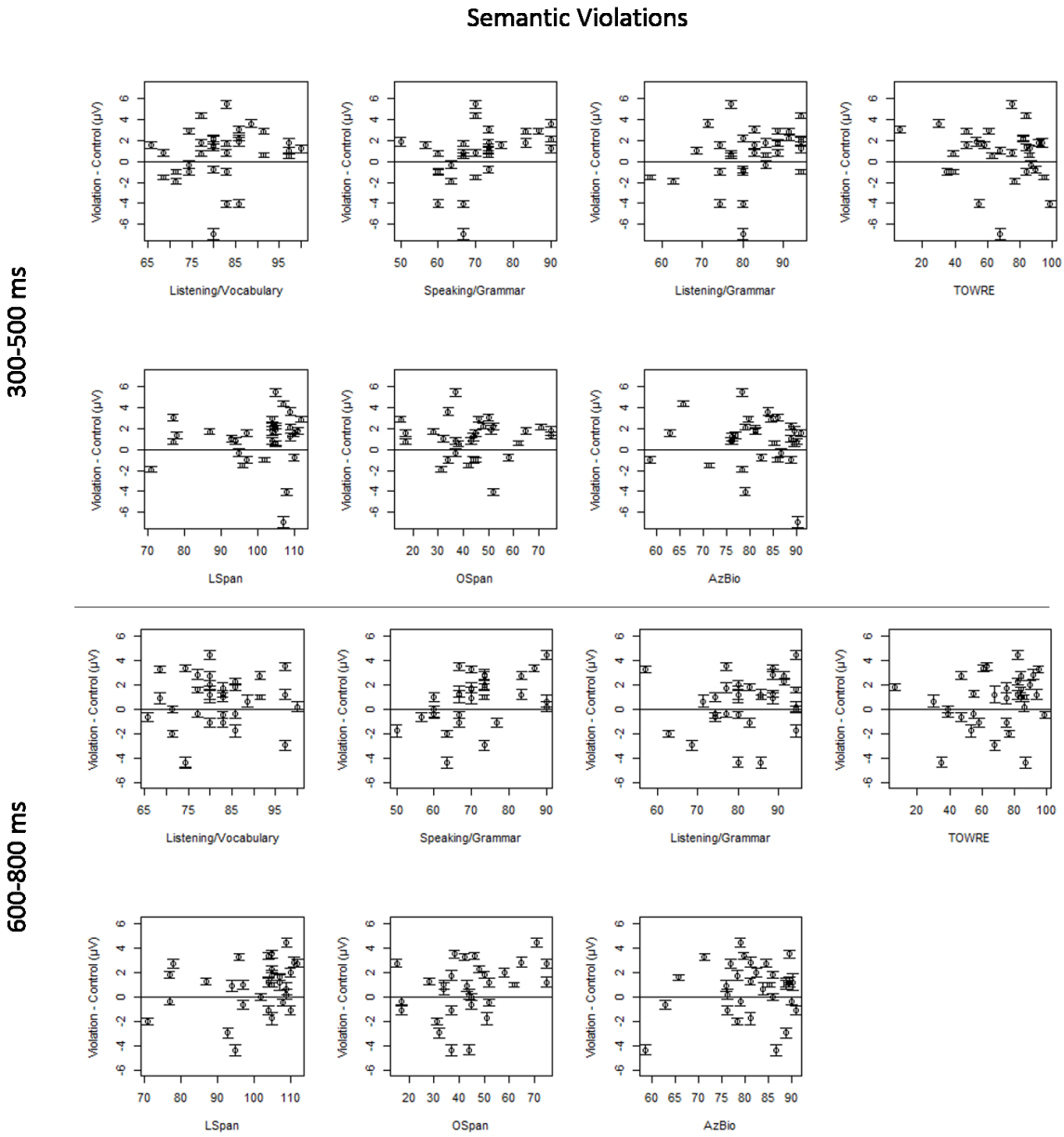


**Figure 3.4** Participants' scores on ID measures in relation to the violation effect (violation – control) for semantic violations in each of the two time windows. Observations were limited to the central midline ROI. Error bars indicate the standard error of observations for each participant.

The overall amplitude of responses to semantic violations during the 300-500 ms time window was found to be negative across the majority of participants, in accordance with the negative deflection in the time course during this time period seen in **Figure 3.1**, and the central parietal negativity during this time window seen in **Figure 3.2**. However, while the mean

amplitude across participants appeared negative, not all participants demonstrated an overall negative response. A minority of participants showed an overall positive response during this time window. Importantly, the results for individuals across ID measures have been averaged across electrodes, and so variation in the topography of responses cannot be ascertained from **Figure 3.4** and **Figure 3.5**. Therefore, an overall mean amplitude of zero μV might arise from an equal distribution of electrodes showing a positive and negative response, and moreover an overall positive response might still be found in participants who showed a negative response at some scalp regions. Rather than providing a detailed description of the effects of ID measures on scalp topography, these figures served primarily to provide a general overview of the impact that ID measures may demonstrate on response amplitude.

Again, in accordance with the time course shown in **Figure 3.1** and the topography in **Figure 3.2**, the majority of participants showed an overall positive response to semantic violations in the 600-800 ms time window. Similarly to the earlier time window, a minority of participants either showed an overall violation effect that opposed this trend, but again, differences in scalp topography which will play a role in understanding these differences likely contributed to these averages in nuanced ways, which will be investigated more deeply in the following chapters.

In general, a number of ID measures appeared to be associated with response amplitude, even without considering variations in scalp topography. For example, the amplitude of responses during the 300-500 ms time window appeared to be strongest in participants with lower scores in the OSpan, Listening/Grammar, Listening/Vocabulary, and Speaking/Grammar tasks. In particular, while participants with higher Speaking/Grammar scores showed little variability in response amplitude, the largest differences appeared to be between participants with lower scores, suggesting that this relationship may be best-modeled using nonlinear functions. Conversely, participants with higher TOWRE scores appeared to demonstrate a

stronger positive response during the 600-800 ms time window. These effects will be investigated in more detail later using the analytical approaches which have been outlined above.

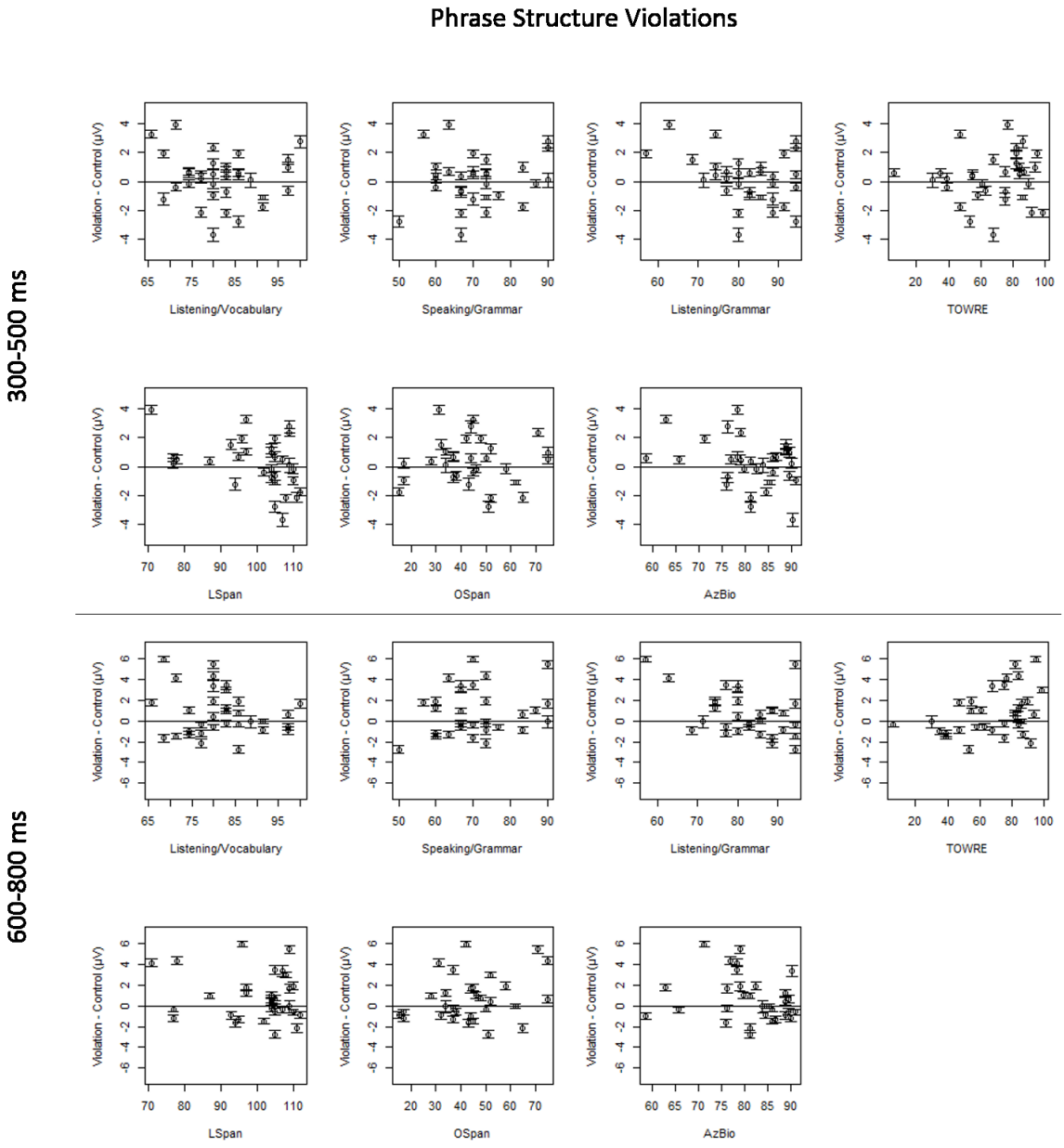


**Figure 3.5** Participants' scores on ID measures in relation to the violation effect (violation – control) for phrase structure violations in each of the two time windows. Observations were limited to the central midline ROI. Error bars indicate the standard error of observations for each participant.

The overall responses to phrase structure violations during the 300-500 ms time window appeared much lower in amplitude, largely centering on zero μV, as seen in **Figure 3.5**. These findings were consistent with the time course of the violation effect during this time window shown in **Figure 3.1**, and the topography shown in **Figure 3.2**, which suggested only a modest effect size which was positive in some regions and negative in others. Moreover, this response amplitude did not appear to be strongly related to any of the ID measures, with the possible exception that negative responses were limited to participants with higher LSpan and Listening/Grammar scores.

Responses to phrase structure violations during the 600-800 ms time window generally suggested an overall positive response across participants, with a minority showing weak negative responses. Again, where these responses are averaged across electrodes, responses with an overall negative amplitude may have arisen from a combination of electrodes showing positive and negative responses across different scalp regions, as **Figure 3.2** suggested may be the case. Interestingly, response amplitude was highest for participants with higher OSpan and TOWRE scores, while those with lower scores showed a weaker violation effect or none at all. Response amplitude appeared to vary most strongly in participants with higher TOWRE scores, while those with lower scores showed little variation. Similarly to responses to semantic violations in the 300-500 ms time window, this pattern suggested that nonlinear modeling techniques may be best-suited to describe this type of relationship.

These findings provided compelling evidence that the ID measures which have been used in the present study were likely related to differences in response amplitude, and potentially to differences in scalp topography as well. The following chapter will evaluate these effects using an approach which is similar to the multivariate regression technique used in past studies in order to describe the relationships between these ID measures and response amplitude in

greater detail, while providing a baseline analysis against which less common techniques will be compared in later chapters (Tanner et al., 2014; Tanner & Van Hell, 2014).

# Chapter 4: Linear mixed modeling of language proficiency and cognition in language violation processing

## 4.1. Introduction

In moving toward an understanding of the link between language proficiency/cognition and language processing, we must consider improvements that can be made in approaches to defining these relationships. Much of the research that has begun to characterize the influence of IDs on language-related ERP components has relied on either ANOVA (Moreno & Kutas, 2005; Pakulak & Neville, 2010) or linear regression models which include only fixed effects (Tanner et al., 2014; Tanner & Van Hell, 2014). In accordance with this research, this chapter has two aims. First, we aim to produce an analysis that can be used as a baseline against which subsequent investigations can be compared. This baseline was predominantly modeled after the best practices that have been outlined in the studies addressed above. However, a second aim was to incorporate a number of analytical developments, which have seldom been used in this area but might yield improved results. This includes testing the impact of moving to mixed modeling procedures, which have been successfully used to describe the influence of ID measures on language-related ERP component characteristics in the past (Newman et al., 2012).

All of the studies discussed in Chapter 1 have described the influences of IDs using the general linear model (GLM), and its various implementations. GLM is a widespread tool in ERP research due to its powerful but easy-to-use implementations, which are native to popular analysis software including SPSS and R. At its core, a Gaussian GLM (such as those used in the present analyses) operates by estimating mean differences in a predicted response (dependent variable) such as scalp voltage, based on independent variables described by the user. These can

include categorical variables (or factors) such as sex or treatment condition, or continuous ones such as working memory capacity. An estimate is considered reliable if the variance in the dependent variable that it explains (i.e., the magnitude of the estimate) significantly exceeds the residual, or unexplained, variance (Sánchez, 1982).

Notably, however, this simplicity comes with several assumptions, and if these assumptions are not met then the predictions of the model may be invalid. Therefore, these assumptions can also be considered limitations in the generalizability of GLM under non-ideal circumstances. First, by default GLM assumes that a linear relationship exists between changes in a continuous independent variable and the dependent variable. Problems with this assumption typically become evident through non-normally distributed residuals. Notably, non-normally distributed residuals can also arise from inconsistent variance across an independent variable. In either case, the result can be invalid estimates of effect size, and therefore an incorrect evaluation of significance, depending on the degree to which a nonlinear relationship exists in the data. Such findings can also be problematic for interpretation as it may suggest the presence of an effect where one does not exist, or conversely that no effect exists when the opposite might be true. Some instantiations of GLM allow the user to define nonlinear (e.g., exponential functions), and provided these accurately depict observations, these concerns can be mitigated (Jiang, 2007). These approaches will be considered in greater depth in Chapter 4.

An additional characteristic of a Gaussian GLM, which was the GLM family used in the present analyses, is that it assumes homoscedasticity. This requires that the degree of residual variance is equal between factor groupings, or along the spectrum of a continuous dependent variable. This results from the assumption that sources of error are not correlated with one another and are uniformly distributed. If this assumption is not satisfied, it can invalidate hypothesis testing. No real-world data satisfies this assumption completely, but a degree of

deviance is tolerable for GLM. Therefore any violations of this assumption should be tested prior to interpretation of a model's findings. Lastly, the influence of observations on an estimate should be equal. That is, no observation should have considerably more 'push' or 'pull' on the direction of an effect than others. This can happen if extreme values of the dependent variable occur (i.e., outliers). Again, while real-world data rarely adheres completely, the degree to which this assumption is satisfied can be assessed by calculating leverage (Cardinali, 2013) and Cook's distance (Cook, 1977).

While violations of the above assumptions can be problematic for GLM, several measures exist to alleviate their ramifications. As discussed in Chapter 1, one way of improving model accuracy and generalizability is through the inclusion of complex *random effects* structures to adjust for sources of variance that are known (explainable), but nonetheless not of direct relevance to the research question. For example, while a predicted outcome may follow administration of a drug, individuals can be expected to vary in the response to the drug. If only variance associated with the experimental variables is accounted for in the model, this between-participant variance will increase unexplained variance of the model, at the expense of significance of the treatment effect. On the other hand, if the variance associated with the random sampling of participants from the population is accounted for in the model, the proportion of unexplained variance is reduced, increasing sensitivity to the experimental effects. In this sense, reporting a 'maximal' random effect structure can be one approach to developing an exhaustively-descriptive model with the aim to improve generalizability and sensitivity (Barr et al., 2013).

Allowing for random variation to be deemed 'explained variance' is also likely to result in reduced residuals with improved normality, and so it is important to do so only where its source is justifiable and cannot be predicted as a fixed effect. Otherwise, reckless inclusion of random

effects can improve apparent fit of the model to any data set by describing variance that may not generalize beyond the sample, bolstering significance and diminishing applicability to future random samples. In addition, models which include random effect structures that are inordinately complex may be uninterpretable, or fail to converge altogether if the underlying data set is not large or diverse enough to support the complexity of the effects (D Bates, Kliegl, Vasishth, & Baayen, 2015). Nonetheless, when used properly, random effects can alleviate violations of GLM's assumptions, improve detection of subtle effects, and benefit validity and reliability.

The specification of random effects as described above, as well as identifying which fixed effects should be included in a model, together comprise the model selection process. A number of issues surrounding the model selection process were outlined in Chapter 1. Most notably, this includes specifying a random effect structure that is appropriate for the experimental design and hypotheses, and avoiding multicollinearity among fixed effects. Regarding fixed effects specifically, the steps to arrive at the ideal model are not always clear, given that a variety of ID measures may be available and only a fraction of those may be useful in predicting response amplitude. Moreover, including predictors which are correlated with one another in a model can be harmful to estimation of their individual effects, which will be an important consideration in the present study. Therefore, it will be important to develop a framework whereby collinear predictor variables are identified and eliminated. In terms of identifying collinear predictors, correlations among all pairs of predictors were evaluated to reveal those which were most associated with one another. Following this, their individual contributions to a model of ERP response amplitude across regions were evaluated to identify those which maximized the model's likelihood in isolation of all other predictors. This assessment relied on the Akaike Information Criterion, which optimizes for a minimum number of model terms and maximal log-

likelihood (Akaike, 1974). In our case, where these comparisons between predictor variables hold the number of observations and terms constant, likelihood was optimized purely by the quality of the predictor.

It should be noted that alternative approaches to addressing multicollinearity exist, such as using principal component analysis to create a single predictor which represents a combination of predictors which had been correlated (Hotelling, 1933; Pearson, 1901). While this approach avoids problems associated with numerous related predictors, interpretation of effects that are specific to individual predictors is made considerably more complex and requires additional analysis (e.g., partial least squares analysis). While the above discussions only touch on the potential impact these decisions can have on results, the present investigation will explore their outcomes in fuller detail. The validity of common techniques in language ERP analysis will be tested to determine ideal methodology, and established effects of individual differences will be validated and expanded upon where possible. In accordance with the concerns addressed above, we examined and attempted to optimize the model-building process as it relates to language violation processing, and its interaction with ID measures across several domains. To achieve this, we used ERP and ID measure data as described in Chapter 2 to identify domains which may predict cortical processing of language violations.

This analysis was intended in part to provide a baseline modeling procedure against which less common techniques could be compared in subsequent chapters. Methodologically, the current investigation included two areas of focus. The first was to quantify the improvements that resulted from including random effects, both in terms of participant-specific baseline responses and topographical distributions. A bare-bones linear model which included only fixed effects was compared with several variations of linear mixed models, which differed only in their specification of random effects, to characterize the impact of their inclusion. We

then evaluated the likelihood that each model could have produced the experimental data (log-likelihood) in order to quantify any improvement that specific random effect structures could provide to the models.

Our second focus was to develop a framework for selecting an ideal set of ID measures, as it was unknown which ones may prove interesting/meaningful, but nonetheless a deliberate and relatively small set must be selected in order to avoid multicollinearity. Given the number of ID measures that were chosen for investigation, the model selection process was optimized through consideration of correlations between measures and the validity of their inclusion as model terms, where the goal was to arrive at the most descriptive but parsimonious model possible. This selection process again used the log-likelihood of the model as a measure of improvement, but with penalties for inclusion of additional model terms, where a balance between the two should yield the ideal model (Akaike, 1974).

It was hypothesized that the expense of including random effects (both participant-specific baseline responses and topographies) would be outweighed by the benefit of their inclusion (in terms of model likelihood). This was expected to result in a significant improvement in model likelihood that was associated with including random mean response amplitudes for each participant, as well as allowing random variance in response topography between participants. Furthermore, it was hypothesized that the ideal model would not include all ID measures as predictors due to either multicollinearity or minimal predictive power from some, resulting in a focus on only the most salient ID measures. It should be noted that the present study includes more unique ID measures than any previous work in this area to our knowledge, and so it was not known which would serve as the strongest predictors. Rather, these findings served as an exploratory introduction to gauge the efficacy of these ID measures, and to provide a framework for comparison with subsequent chapters.

## 4.2. Methods

All details pertaining to the collection and pre-processing of data are as described in Chapter 2, *Data Collection and Pre-Processing*. The following methods were specific to the present investigation.

### 4.2.1. Random Effect Selection

Selecting the appropriate random effect structure for an experimental design can have a considerable impact on the estimates (coefficients) of fixed effects. When used appropriately, the result can be reduced error in the model and improved predictive accuracy ($R^2$), and by extension results that generalize beyond the experimental setting more accurately than those obtained using fixed effects only. This is due to allowing flexibility in the calculation of terms where unknowable variability is expected. For example, where language proficiency might be predicted to have an overall effect on cortical response amplitude, two unknowable factors may confound the results: 1) individual participants will likely have different baseline responses, against which changes must be considered, and 2) the topographical distribution of effects may differ between individuals. In the above example, a model containing no random effects is derived using the following formula:

*Amplitude ~ Proficiency*

In this formula, *Amplitude* is cortical response (μV), and *Proficiency* is a continuous predictor indicating some aspect of language proficiency for a participant. Any variability in the base response between participants, and any variability in the topography of responses, is not modeled, and becomes a part of the error term. This can be improved upon as following:

*Amplitude ~ Proficiency + (1|Participant)*

In this formula, *Participant* represents a factor with a unique mean for each participant. The result is a random base response for each participant, allowing flexibility in the intercept for each participant. This effect will be referred to as a random by-participant intercept, referring to the mean response amplitude across trials, conditions and electrodes for each participant. However, there is still no accounting of variability in the scalp topography across participants. The following achieves this:

*Amplitude ~ Proficiency + (1 + ROI | Participant)*

In this formula, a random effect of scalp region, *ROI*, has been nested within the random by-participant intercept. It therefore predicts that a degree of unknowable variance in the topography will exist within each participant. This effect will be referred to as a random ROI-by-participant slope. Note this formula allows for a degree of correlation between the random ROI-by-participant slope and by-participant intercept. That is, there are no constraints placed on the estimates, and correlation between the two is expected to reflect the observed data. If there is a requirement to force zero correlation between the two, which might produce more orthogonal random effect estimates but make achieving a better fit more difficult, then the formula can be amended as follows:

*Amplitude ~ Proficiency + (1 | Participant) + (0 + ROI | Participant)*

This type of random effect structure was not expected to result in any benefit to the present study, though it was investigated as an exploratory measure. However, an experimental design that must model two potentially related terms as random effects should consider the degree of relationship between those two variables when deciding on a random effect structure, as non-unique dependency of a response on multiple predictors (multicollinearity) is detrimental to model viability in fixed and random effects alike. Either of the previous two models will be a

candidate for depicting random variability between participants, both in base response amplitude and scalp distribution. Therefore, we investigated the capacity for each of the above random effect structures to maximize the descriptive ability of the resulting model.

### 4.2.2. Linear Modeling

The effects of language proficiency and cognition were evaluated using linear mixed effects modeling, using the *lme4* package (Bates et al., 2011) with *R version 3.4.1 (R Development Core Team, 2013)*. As described in Chapter 2, predictor variables of interest included Listening/Vocabulary, Speaking/Grammar, Listening/Grammar (Hammill et al., 1994), word reading efficiency (TOWRE-2), two measures of working memory capacity (OSpan and LSpan), and speech comprehension (AzBio). Scalp voltage was averaged across each time window of interest, including 300-500 ms and 600-800 ms following the onset of the violating word, for each electrode.

Multicollinearity is known to be problematic for the general linear model. In extreme cases, two strongly correlated predictors are nearly equally capable of explaining variance in the dependent variable, but are incapable being assigned coefficients which predict that same variance. Their effect size is necessarily reduced, in proportion to the strength of their correlation, and distinctions between the predictors cannot be made. This is true of any additive modeling solution (e.g., GLM), as model terms (along with residuals) sum to a predicted response, and depicting separate correlated predictors as equally capable of predicting a response would overestimate its magnitude of the prediction. Therefore, perfectly orthogonal (uncorrelated) predictors are ideal, but this is rarely possible. Instead, measures must be taken to ensure that collinearity between predictors is minimized where possible.

Collinearity of our cognitive and language proficiency measures was assessed using Spearman's $\rho^2$, where only one predictor was selected from any group of related measures for potential contribution to the final model. This determination was made using the Akaike Information Criterion (AIC; Akaike, 1974), a measure which applies penalties for addition of terms against the log-likelihood of a model. The AIC is calculated as shown below in Formula 1, where *k* is the number of model terms, and *L* is the log-likelihood of the resulting model. The outcomes of this analysis will be discussed in more detail in the results below.

$$AIC = 2k - 2\ln(L) \qquad (1)$$

As each collinear predictor was evaluated independently, the number of terms in the models being compared was unchanged. Therefore, use of the AIC in this instance was equivalent to evaluating changes in the log-likelihood of models through inclusion of individual predictors. This resulted in removal of predictors which were highly correlated, but less likely to have produced the observed data.

The modelling process was completed in four separate instances, one for each of the two time windows (300-500 ms and 600-800 ms) and two sentence types (semantic and phrase structure) being investigated. Potential predictors in each model were determined using the same process, but since each outcome measure was associated with different eliciting conditions (violation type) and underlying cognitive processes, it was possible for different predictors to be included across models.

For each model, we first created a linear model that included scalp voltage as a function of fixed effects for sentence condition (control or violation), ROI, and their interaction. This was compared with a model that included these same fixed effects as well as a random intercept for each participant, and a random ROI effect for each participant. As above, the model with the

lower AIC value was taken as that which best predicted our observations. Next, the inclusion of model terms for participant age and sex was tested in the same manner. All models were improved by the addition of the described random effects, but none benefited from the inclusion of age or sex. This resulted in our base model (Formula 2):

$$Amplitude \sim condition \times ROI + (1 + ROI|participant) \quad (2)$$

It was expected that participants would vary in their overall voltage output at the scalp, and that depiction of effects across the scalp would differ between participants based on unknowable differences in scalp tissue, head shape, and electrophysiology. An investigation into the optimal random effect structure is detailed below. While random inconsistencies between the violation effects of different sentence types may also be expected between participants, and the distribution of those differences may vary unknowably across the scalp as well, any given model predicted responses for only a single sentence type. This precludes the requirement for random sentence type effects across participants. Each sentence type and time window were modeled separately to aid interpretability by limiting the violation effect to only a three-way interaction (i.e., *condition × ROI × score*) rather than four- or five-way interactions (i.e., *sentence × time × condition × ROI × score),* despite that modeling subsets of the full data set was likely related to a reduction in statistical power.

As described above, only a subset of predictors could be included in a model, where those included were based on the probability that a predictor would have yielded the observed scalp effects (model likelihood). Consequently, allowing for each model to include potentially non-overlapping predictors from correlated sets avoided the assumption that there was one 'best' set of predictors for all sentence types and time windows. Pairwise correlations were used to address collinearity among predictors, and no model included any two or more collinear pairs of ID measures (Spearman's $\rho^2 > 0.1$). While these results are described in greater detail in the

results, the ID measure groupings are pertinent to the model building process and so they are briefly described here. Collinear pairs of variables included: 1) TOAL-3 Listening/Vocabulary and AzBio, 2) TOWRE-2 and OSpan, and 3) TOAL-3 Speaking/Grammar and LSpan. TOAL-3 Listening/Grammar was not strongly correlated with any other predictors and was investigated in every model.

#### 4.2.3. Model Selection

In each of the four models, an ideal model was created in a stepwise fashion as follows: Each potential predictor, as determined above, was included in a new variant of the base model, which allowed for interactions between condition (violation or control), ROI, and the predictor. This resulted in as many models as there were predictors, and the model that produced the lowest AIC became the new base model against which the addition of each remaining potential predictors was evaluated. The model continued to be built using this procedure until either inclusion of any remaining predictors did not represent a substantial improvement in the model, or all predictors were included. The determination of a "substantial" improvement in this case was based on an AIC improvement of five or greater, as an improvement of less than five is generally considered not to provide meaningful support for a model (K P Burnham & Anderson, 2004).

Importantly, the AIC does not allow for determination of significance, but is intended to suggest favorability when presented with two or more models. AIC values are strongly related to sample size, which can make comparisons ambiguous in models that consider differing numbers of observations. Attempts have been made to improve interpretability of raw AIC values, for example through transformation into Akaike weights (Wagenmakers & Farrell, 2004). While Akaike weights do allow for comparison of likelihood between models that include different

sample sizes, the determination of a likelihood threshold is still required and is no less ambiguous than comparisons using raw AIC values alone. However, because the sample sizes were identical in all comparisons here, raw AIC values were considered to be a direct indicator of model likelihood.

ID measures were included such that up to three-way interactions between the measure, condition, and ROI were allowed, but that ID measures did not interact with one another. This level of interaction was chosen to aid interpretability; changes in violation effect size as a function of an ID measure for an ROI were central to our investigation, but interactions between ID measures were not. In addition, higher-order interactions (e.g., effects that are only applicable to specific combinations of numerous ID measure scores) have limited generalizability and are cumbersome to interpret. Once measures were chosen, the final model was determined through iteratively removing individual fixed effects—starting with highest-order effects (in this case, three-way interactions)—that had *F* statistics that corresponded with *p* values above alpha. Non-significant lower-order terms were kept if they were part of a higher-order interaction. This step redefines variance explained by non-significant terms as model residuals, further reducing Type I error. This method of iterative back-fitting was performed using the *LMERConvenienceFunctions* package (Tremblay & Ransijn, 2013) in *R (R Development Core Team, 2013)*.

The significance of interactions between an ID measure and the violation vs. control sentence difference term was evaluated at each ROI using the three-way interaction between condition, ROI, and an ID measure, with a Bonferroni adjustment made for the 9 ROIs being investigated. In many cases, interactions were significant (suggesting that the violation effect changes as a function of the ID measure), but effect sizes were small. Therefore, there was an additional requirement that the 95% confidence interval of the violation - control difference did

not contain zero for some portion of the ID measure spectrum. This ensured that the ERP amplitudes for violation and control sentences diverged to a considerable degree (i.e., responses to each of the two conditions differed in amplitude) for the individuals in that portion of the ID measure spectrum. If both requirements were met, such that the ID measure significantly influenced the conditional contrast (as indicated by the significance of the interaction term), and the conditional contrast was significantly greater than 0 μV, the relationship between an ID measure and the violation effect was considered significant at an ROI.

Investigations of any ID measure only included individuals with scores that fell within the range of the mean ± 3 standard deviations for that measure. This resulted in removal of one individual from investigations that included the AzBio speech comprehension measure.

## 4.3. Results

### 4.3.1. Data Quality and Random Effects

Prior to evaluating the results of any models, the quality of the data and the impact of including random effects in a model was evaluated. **Figure 4.1** shows the distribution of scalp voltage observed for each time range of interest (300-500 ms and 600-800 ms). Responses demonstrated a relatively normal distribution in each time range.

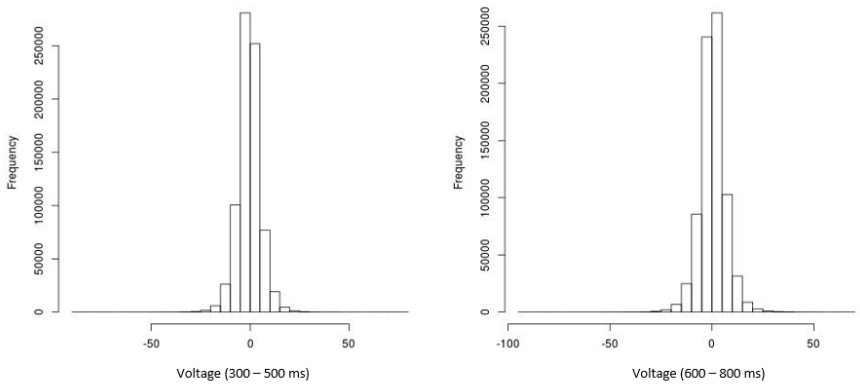


**Figure 4.1** Scalp voltage frequency each time range of interest: 300-500 ms (left), and 600-800 ms (right), including both semantic and phrase structure violation responses.

These responses were fit to two models: 1) a simple GLM using only fixed effects to evaluate scalp voltage following onset of the violating word as a function of condition (violation vs. control), sentence type (semantic or phrase structure), and ROI, and 2) a similar model, fit using LME, which was identical except for the addition of a random intercept for each ROI in each participant. These models were fit in order to assess the improvements afforded through inclusion of random effects. Model residuals for the two approaches are shown in **Figure 4.2**, qualitatively suggesting an improvement in model sensitivity (i.e., reduced noise) for LME. The scales of values predicted by each model (*x*-axis) differed slightly, as inclusion of random effects resulted in predicting stronger signals at times. However, the magnitude of residuals (*y*-axis) was ostensibly similar between the two models, with a maximum of approximately 50 μV for any given predicted value in each.

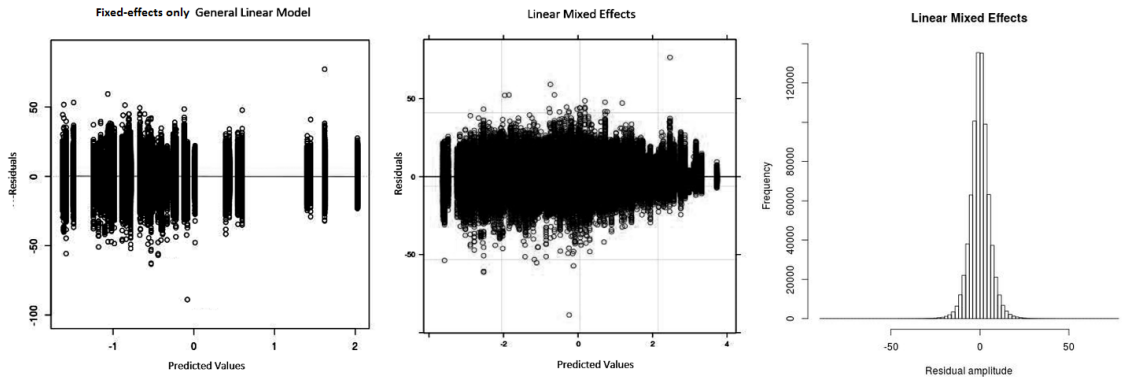


***Figure 4.2*** *Residuals vs model-predicted values using the fixed-effects only general linear model (left), and linear mixed effects model (middle). Normality of residuals is demonstrated for the linear mixed effects model (right).*

Addition of this random effect improved the AIC significantly (by 16,241, where 5 is considered significant; Akaike, 1974), reflecting greater model likelihood in the mixed model. Notably, including random effects was associated with decreasing variance as predicted values increased. Residuals were also normally distributed as seen in the right-most pane of **Figure 4.2**. To further evaluate normality, a Q-Q plot for the above linear mixed model is shown in **Figure 4.3**. While there was some deviation of observed quantiles from the theoretical normal distribution at either end (i.e., heavy tails), this was not so strong as to consider the model invalid. Nonetheless, this suggests that non-parametric hypothesis testing methods may improve model accuracy above the parametric GLM approaches used in this research.

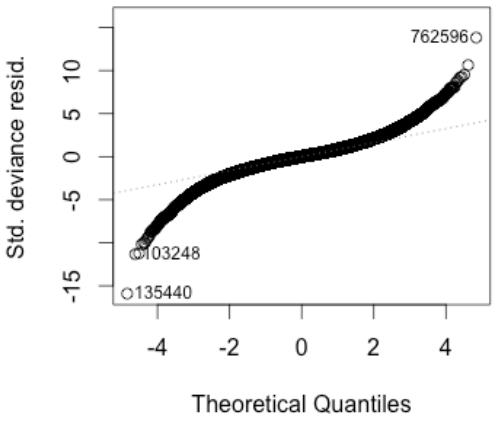


**Figure 4.3** Q-Q plot for a linear mixed model including fixed terms for sentence type, condition, ROI, and random terms for participant and ROI within participants. Deviation of quantiles from the theoretical normal distribution on either end suggests a degree of non-normality in the distribution.

Regarding homogeneity of variance in model terms, formal tests (e.g., Bartlett's or Levene's test for heteroscedasticity) are heavily dependent on sample size, and given that the depicted models were built on more than 770,000 observations, even marginal differences in residual variance between factor levels could be considered significant violations. One possible solution might be to downsample the data set by testing only a randomized subset of observations, arriving at a data set size that is appropriate for testing. However, this risks excluding problematic observations which may exist. Therefore, these tests cannot be relied on in the present sample. However, in all cases, residuals for each factor level demonstrated a mean of ostensibly zero with similar variance (e.g., $2.3 \times 10^{-15} \pm SD = 5.55$ µV for control sentences, and $-1.7 \times 10^{-14} \pm SD = 5.54$ µV for violation sentences), and so heterogeneity of variance was not considered a concern.

### 4.3.2. Predictor Collinearity

As discussed, correlation between observations in distinct predictor variables (i.e., collinearity) is problematic for the general linear model. Correlation of predictors in a model can

result errors in effect size estimation (Barr et al., 2013). Therefore, it was necessary to address collinearity, and for any set of correlated variables include only those that most improved a model. Spearman's $\rho^2$ was calculated between all sets of predictors, and any two or more of sufficient similarity were never included in the same model. Here, sufficient similarity was defined as $\rho^2 > 0.1$, a conservative estimate which was intended to select for only the most orthogonal ID measures. **Figure 4.4** describes the collinearity of all predictors. Those found to be similar were (1) Listening/Vocabulary and AzBio, (2) TOWRE and OSpan, and (3) Speaking/Grammar and LSpan. For each set, only the predictor that most improved the model's description of the variance was used. Because a separate model was created for each sentence type and time window of interest, these models frequently included different predictors, as described below.

Notably, there are alternatives to such iterative solutions for predictor selection. For example, the elastic net method has proven capable in data sets where the number of predictors outnumber the number of observations (Zou & Hastie, 2005). While this method is also applicable for use with fewer predictors, it is known to characteristically include or exclude correlated predictors as a group, and is not suitable for the purpose of eliminating one of a pair of correlated predictors (Zou & Hastie, 2005). Similarly, partial least squares analysis can be used to identify these relationships without the need to fit numerous models, selecting predictors (or groups of predictors) which explain variance in a response variable (Bry, Trottier, Mortier, Cornu, & Verron, 2016). However, this method has similarly failed to select between correlated predictors (Bry et al., 2016), and neither is therefore well-suited to address multicollinearity.

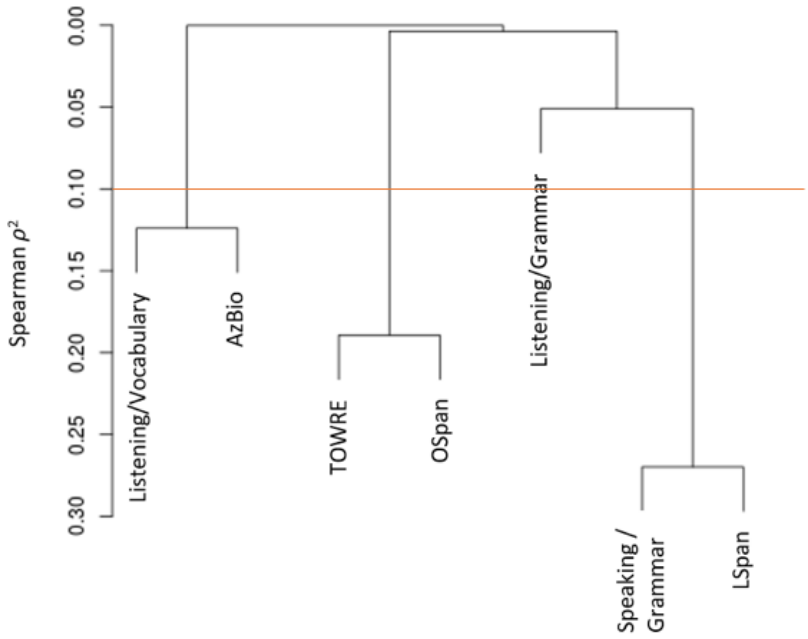


***Figure 4.4*** *Dendrogram of Spearman's ρ² indicating collinearity of predictor variables. Predictors with a correlation beyond the threshold (ρ² > 0.1), shown as a red line, were not included in the same model together.*

### 4.3.3. Random Effect Structure

For this investigation, we evaluated the candidacy of random effects structures primarily using AIC and the related Akaike weights (Wagenmakers & Farrell, 2004). In addition, we considered conditional $R^2$. We used conditional $R^2$ as a measure of predictive accuracy (as opposed to marginal $R^2$, which will be of use during later comparisons of time windows and sentence types) because this measure includes random effects when predicting responses, and the impact of random effect inclusion was central to this investigation. **Table 4.1** shows an example of the results of this investigation for phrase structure violations in the 600-800 ms time window.

***Table 4.1*** *Degrees of freedom (DOF), AIC, ΔAIC, Akaike Weight, and conditional $R^2$ for each of four potential random effect structures: 1) None 2) By-participant intercepts, 3) ROI-by- participant slopes with by-participant intercepts, and 4) ROI-by- participant slopes with by- participant intercepts at forced zero correlation. Models were created for phrase structure violations in the 600-800 ms time window.*

| RANDOM EFFECTS | DOF | AIC | AKAIKE WEIGHT | CONDITIONAL $R^2$ |
|---|---|---|---|---|
| **NONE** | 91 | 1,323,814 | < 0.001 | 2.41% |
| **(1\|PARTICIPANT)** | 92 | 1,321,852 | < 0.001 | 3.77% |
| **(1+ROI\| PARTICIPANT)** | 136 | 1,320,468 | 0.874 | 4.58% |
| **(1\| PARTICIPANT) + (0+ROI\| PARTICIPANT)** | 137 | 1,320,472 | 0.126 | 4.58% |

In this investigation, using region-by-participant slopes with by-participant intercepts represented the optimal random effect structure. This model produces the most favorable AIC to a considerable degree when compared with the previous two (simpler) models, which was our primary metric of model parsimony and quality. The added constraint of forcing zero correlation between the two effects resulted in a more complex but not substantially-improved model. This was also reflected in the Akaike weights, which favored this model to a considerable degree. In addition, the predictive accuracy (conditional $R^2$) is strongest for either model that included ROI as a random effect. Again, opting for the simpler model encourages use of the former. Therefore, inclusion of random region-by-participant slopes and by-participant intercepts produces the most parsimonious and accurate model. This random effect structure was therefore used in each time window and for each sentence type.

### 4.3.4. Semantic Violations, 300-500 ms

The result of AIC-driven model for semantic violations in 300-500 ms is shown in **Table 4.2**. Post-hoc testing of those terms in bold is shown in **Table 4.3**.

***Table 4.2*** *Model terms for responses 300-500 ms following the onset of semantic violations. Significance of a term is denoted using* * ($p < .05$), ** ($p < .01$), or *** ($p < .001$).

| | ***F*** | ***DOF*** | ***p*** | |
|---|---|---|---|---|
| *Condition* | 762.26 | 1, 201205 | 0.000 | *** |
| *ROI* | 7.42 | 8, 201205 | 0.000 | *** |
| *OSpan* | 1.07 | 1, 201205 | 0.301 | |
| *Speaking/Grammar* | 0.04 | 1, 201205 | 0.835 | |
| *Listening/Grammar* | 0.04 | 1, 201205 | 0.850 | |
| ***Condition:ROI*** | 351.40 | 8, 201205 | 0.000 | *** |
| *Condition:OSpan* | 20.07 | 1, 201205 | 0.000 | *** |
| *Condition:Speaking/Grammar* | 642.26 | 1, 201205 | 0.000 | *** |
| *ROI:Speaking/Grammar* | 1.23 | 8, 201205 | 0.279 | |
| *Condition:Listening/Grammar* | 21.12 | 1, 201205 | 0.000 | *** |
| *ROI:Listening/Grammar* | 0.91 | 8, 201205 | 0.506 | |
| ***Condition:ROI:Speaking/Grammar*** | 7.48 | 8, 201205 | 0.000 | *** |
| ***Condition:ROI:Listening/Grammar*** | 20.53 | 8, 201205 | 0.000 | *** |

Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at each ROI. Significance in this contrast was corrected for 9 comparisons using a Bonferroni adjustment. ID measures that were included in three-way interactions with condition and ROI indicate those that were likely to affect violation processing. These measures are denoted in bold and will be discussed below.

The latency and distribution of the negative violation effect seen in the 300-500 ms time window was consistent with the predicted N400 component. As seen in **Table 4.3**, the three-way interaction between Listening/Grammar score, condition, and ROI reflected that the effect of Listening/Grammar score on N400 amplitude was strongest at the anterior midline ROI, where

participants with lower scores showed a larger N400 effect. Listening/Grammar also influenced N400 amplitude at a number of posterior and central ROIs, as outlined in **Table 4.3**. However, for brevity, only ROIs at which the strongest effect of an ID measure are outlined in this text. Similarly, participants with lower Speaking/Grammar scores showed a higher-amplitude N400 across frontal and midline ROIs, with this effect being strongest in the right anterior and right midline ROIs. While the influence of Speaking/Grammar on N400 amplitude was strongest in the left posterior ROI, as seen in the slope for this region in **Table 4.3**, response amplitude at this ROI was lower. Rather, it was the anterior midline ROI which was associated with the highest-amplitude N400. These effects are outlined in **Figure 4.5**.

To elucidate whether the effect of steeper Listening/Grammar slopes at anterior than central ROIs might relate to differences in the topography of the response between lower- and higher-scoring individuals, the response topography was calculated for these two groups, as depicted in **Figure 4.6**. Participants were separated using a median division of Listening/Grammar scores into two quantiles. Lower-scoring participants demonstrated farther-reaching central parietal negativity during this time window than did higher-scoring participants, who instead showed stronger anterior positivity.

It is important to note that, as described above, effects are only depicted as significant if two criteria are met. First, the confidence interval for an estimate (violation vs. control contrast) must not contain zero for some value of an ID measure. Second, the slope of the estimate must be statistically significant. In some cases, ID measures are shown to have a significant effect on response amplitude for a region despite that the confidence interval only exceeds zero at one extreme value (e.g., the effect of Speaking/Grammar score at the central midline ROI, **Figure 4.5**), and while these effects meet our criteria for significance, they may not be replicable and are not reported.

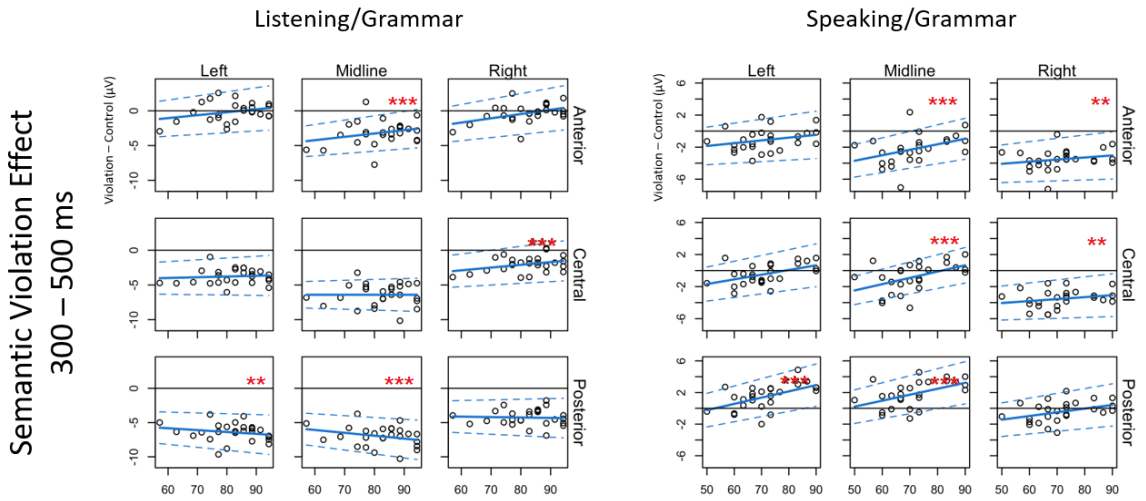


***Figure 4.5*** *Semantic violation effects across each ROI* in the 300-500 ms time window. *The 95% confidence intervals are shown, indicating violation effects, with a significant slope of the predictor shown (** p < .05, *** p < .01, **** p < .001) where the CI of that effect exceeds zero.*

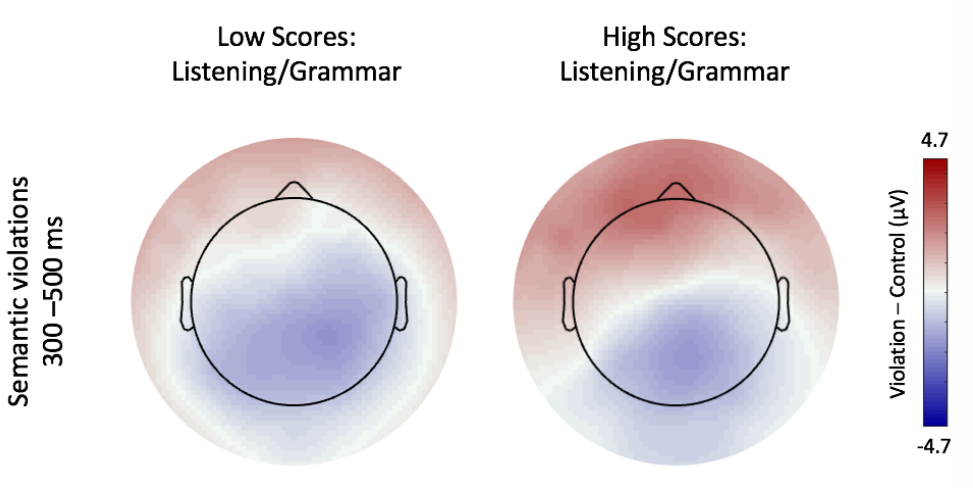


***Figure 4.6*** *Averaged topography of responses to semantic violations 300-500 ms following onset of the violation word for participants below (left) and above (right) the median split for Listening/Grammar scores.*

***Table 4.3*** *Post-hoc comparison of condition (violation vs. well-formed) contrast at each ROI, and three-way interactions (modulation of the condition contrast slope by an ID measure at each ROI) in the 300-500 ms time window for semantic violations. Rows show any ROI where the 95% CI of the condition contrasts exceeded zero, alongside significance of the interaction at that region.*

| Interaction | ROI | Corrected Significance | Model Estimate |
|---|---|---|---|
| Condition x ROI (DOF = 201205) | Anterior Left | t = -4.41, $p < .001$, *** | -3.58 |
| | Anterior Midline | t = -10.14, $p < .001$, *** | -7.13 |
| | Anterior Right | t = -6.63, $p < .001$, *** | -5.38 |
| | Central Left | t = -6.29, $p < .001$, *** | -4.61 |
| | Central Midline | t = -10.53, $p < .001$, *** | -6.41 |
| | Central Right | t = -7.19, $p < .001$, *** | -5.27 |
| | Posterior Left | t = -5.71, $p < .001$, *** | -5.71 |
| | Posterior Midline | t = -4.85, $p < .001$, *** | -3.56 |
| | Posterior Right | t = -5.14, $p < .001$, *** | -3.77 |
| Condition x ROI x Speaking/Grammar (DOF = 201388) | Anterior Midline | t = 10.22, $p < .001$, *** | 0.196 |
| | Anterior Right | t = 3.37, $p = .007$, ** | 0.075 |
| | Central Midline | t = 13.51, $p < .001$, *** | 0.224 |
| | Central Right | t = 3.49, $p = .004$, ** | 0.069 |
| | Posterior Left | t = 11.19, $p < .001$, *** | 0.226 |
| | Posterior Midline | t = 10.62, $p < .001$, *** | 0.216 |
| Condition x ROI x Listening/Grammar (DOF = 201405) | Anterior Midline | t = 6.45, $p < .001$, *** | 0.128 |
| | Central Left | t = 1.33, $p = 1.00$ | 0.028 |
| | Central Midline | t = -0.01, $p = 1.00$ | 0.000 |
| | Central Right | t = 5.08, $p < .001$, *** | 0.106 |
| | Posterior Left | t = 3.50, $p = .004$, ** | 0.073 |
| | Posterior Midline | t = -5.37, $p < .001$, *** | 0.112 |

### 4.3.5. Semantic Violations, 600-800 ms

The result of AIC-driven model for semantic violations in 600-800 ms is shown in **Table 4.4**. Post-hoc testing of those terms in bold is shown in **Table 4.5**.

***Table 4.4*** *Model terms for responses 600-800 ms following the onset of semantic violations. Significance of a term is denoted using * (*p* < .05), ** (*p* < .01), or *** (*p* < .001).*

| | ***F*** | ***DOF*** | ***p*** | |
|---|---|---|---|---|
| *Condition* | 362.97 | 1, 201188 | 0.000 | *** |
| *ROI* | 4.26 | 8, 201188 | 0.000 | *** |
| *OSpan* | 4.74 | 1, 201188 | 0.030 | * |
| *Speaking/Grammar* | 0.00 | 1, 201188 | 0.977 | |
| *Listening/Grammar* | 8.42 | 1, 201188 | 0.004 | ** |
| *AzBio* | 0.01 | 1, 201188 | 0.929 | |
| ***Condition:ROI*** | 132.91 | 8, 201188 | 0.000 | *** |
| *Condition:Speaking/Grammar* | 420.07 | 1, 201188 | 0.000 | *** |
| *ROI:Speaking/Grammar* | 0.74 | 8, 201188 | 0.656 | |
| *Condition:Listening/Grammar* | 7.25 | 1, 201188 | 0.007 | ** |
| *ROI:Listening/Grammar* | 0.97 | 8, 201188 | 0.455 | |
| *Condition:AzBio* | 146.64 | 1, 201188 | 0.000 | *** |
| *ROI:AzBio* | 2.12 | 8, 201188 | 0.030 | * |
| ***Condition:ROI:Speaking/Grammar*** | 13.73 | 8, 201188 | 0.000 | *** |
| ***Condition:ROI:Listening/Grammar*** | 8.21 | 8, 201188 | 0.000 | *** |
| ***Condition:ROI:AzBio*** | 6.55 | 8, 201188 | 0.000 | *** |

Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs except for central left and central right. Significance of this contrast across ROIs was corrected for 9 comparisons using a Bonferroni adjustment. This two-way interaction depicted significant positivity at all posterior ROIs, as well as the central midline ROI, and a negative violation effect at all anterior ROIs. This pattern was reflected in the three-way interactions between Condition, ROI, and each of three ID measures: Speaking/Grammar, Listening/Grammar, and AzBio. These effects are outlined in

**Figure 4.7**. In these interactions, participants with lower Listening/Grammar and Speaking/Grammar scores showed a higher-amplitude positive response at the posterior central and posterior right ROIs than did higher-scoring participants. AzBio did not modulate the response amplitude at posterior ROIs in this time window. Conversely, the negative response seen at anterior ROIs was modulated by each of the three ID measures, but this influence was weak. The significance of these effects is described in **Table 4.5**.

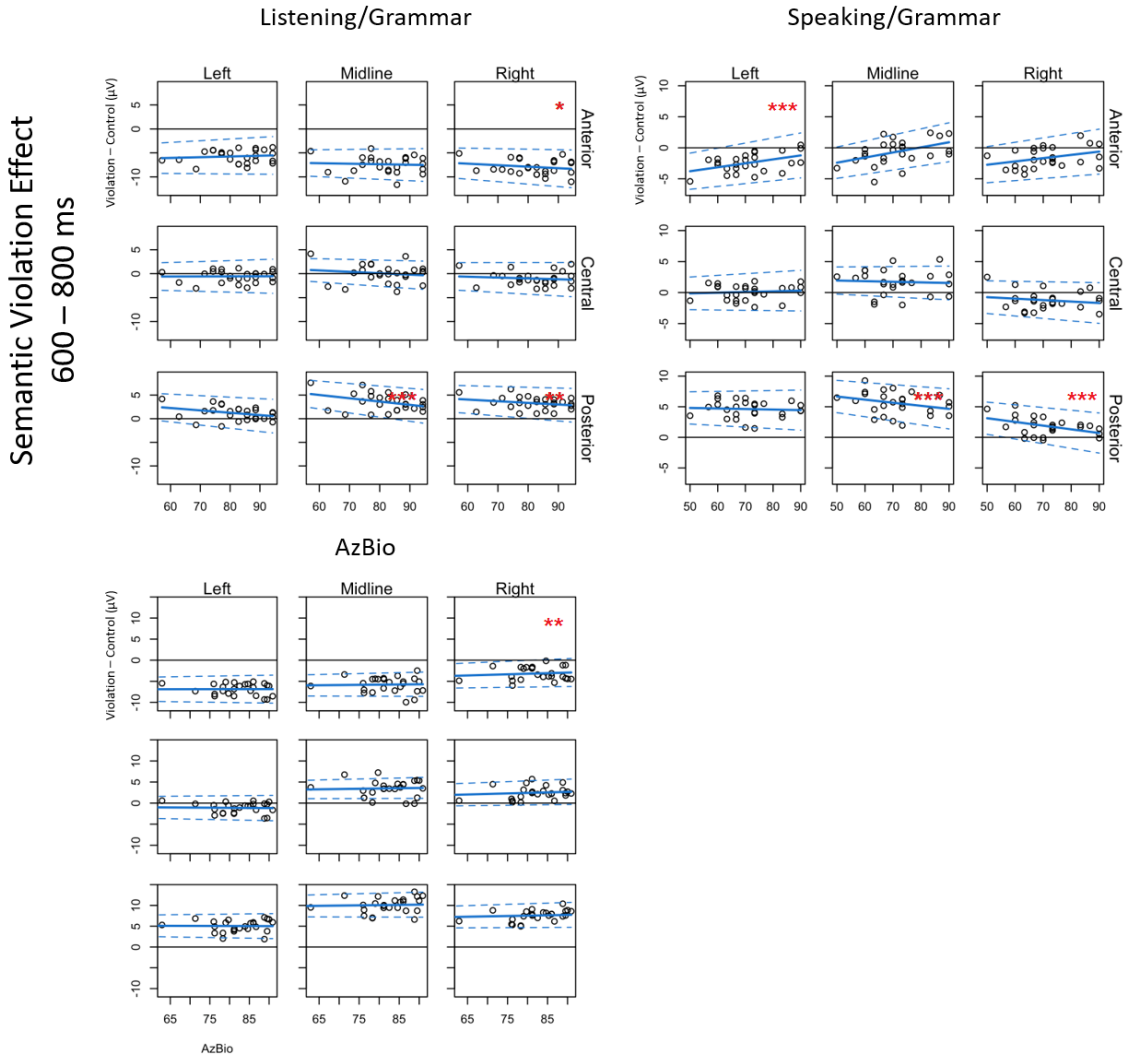


***Figure 4.7*** *Semantic violation effects across each ROI* in the 600-800 ms time window. *The 95% confidence intervals are shown, indicating violation effects, with a significant slope of the predictor shown (** p < .05, *** p < .01, **** p < .001) where the CI of that effect exceeds zero.*

***Table 4.5*** *Post-hoc comparison of condition (violation vs. well-formed) contrast at each ROI, and three-way interactions (modulation of the condition contrast slope by an ID measure at each ROI) in the 600-800 ms time window for semantic violations. Rows show any ROI where the 95% CI of the condition contrasts exceeded zero, alongside significance of the interaction at that region.*

| Interaction | ROI | Corrected Significance | Model Estimate |
|---|---|---|---|
| Condition x ROI (DOF = 201171) | Anterior Left | t = -6.74, $p < .001$, *** | -6.96 |
| | Anterior Midline | t = -7.29, $p < .001$, *** | -6.52 |
| | Anterior Right | t = -5.2, $p < .001$, *** | -5.37 |
| | Central Left | t = -0.72, $p = 1.00$ | -0.67 |
| | Central Midline | t = 3.14, $p = 0.015$, * | 2.43 |
| | Central Right | t = 0.47, $p = 1.00$ | 0.43 |
| | Posterior Left | t = 5.62, $p < .001$, *** | 5.25 |
| | Posterior Midline | t = 9.84, $p < .001$, *** | 9.19 |
| | Posterior Right | t = 6.58, $p < .001$, *** | 6.15 |
| Condition x ROI x Speaking/Grammar (DOF = 201171) | Anterior Left | t = 7.01, $p < .001$, *** | 0.101 |
| | Posterior Left | t = -1.08, $p = 1.00$ | -0.140 |
| | Posterior Midline | t = -6.14, $p < .001$, *** | -0.080 |
| | Posterior Right | t = -7.36, $p < .001$, *** | -0.096 |
| Condition x ROI x Listening/Grammar (DOF = 201171) | Anterior Left | t = 1.47, $p = 1.00$ | 0.021 |
| | Anterior Midline | t = -1.17, $p = 1.00$ | -0.015 |
| | Anterior Right | t = -3.03, $p = .022$, * | -0.044 |
| | Posterior Midline | t = -7.43, $p < .001$, *** | -0.099 |
| | Posterior Right | t = -3.74, $p = .002$, ** | -0.050 |
| Condition x ROI x AzBio (DOF = 201171) | Anterior Left | t = 0.16, $p = 1.00$ | 0.003 |
| | Anterior Midline | t = 1.42, $p = 1.00$ | 0.026 |
| | Anterior Right | t = 3.71, $p = .002$, ** | 0.080 |
| | Central Midline | t = 2.36, $p = .160$ | 0.038 |
| | Posterior Left | t = -0.38, $p = 1.00$ | -0.007 |
| | Posterior Midline | t = 1.65, $p = .877$ | 0.032 |
| | Posterior Right | t = 2.63, $p = .075$ | 0.051 |

### 4.3.6. Phrase Structure Violations, 300-500 ms

The result of AIC-driven model for phrase structure violations in 300-500 ms is shown in **Table 4.6**. Post-hoc testing of those terms in bold is shown in **Table 4.7**.

***Table 4.6*** *Model terms for responses 300-500 ms following the onset of phrase structure violations. Significance of a term is denoted using * ($p < .05$), ** ($p < .01$), or *** ($p < .001$).*

| | ***F*** | ***DOF*** | ***p*** | |
|---|---|---|---|---|
| *Condition* | 6.39 | 1, 205126 | 0.012 | * |
| *ROI* | 9.05 | 8, 205126 | 0.000 | *** |
| *OSpan* | 0.01 | 1, 205126 | 0.920 | |
| *Speaking/Grammar* | 0.55 | 1, 205126 | 0.457 | |
| *Listening/Grammar* | 1.57 | 1, 205126 | 0.210 | |
| *Listening/Vocabulary* | 1.93 | 1, 205126 | 0.164 | |
| ***Condition:ROI*** | 21.75 | 8, 205126 | 0.000 | *** |
| *Condition:OSpan* | 19.18 | 1, 205126 | 0.000 | *** |
| *Condition:Speaking/Grammar* | 3.89 | 1, 205126 | 0.049 | * |
| *ROI:Speaking/Grammar* | 0.45 | 8, 205126 | 0.894 | |
| *Condition:Listening/Grammar* | 307.10 | 1, 205126 | 0.000 | *** |
| *ROI:Listening/Grammar* | 0.91 | 8, 205126 | 0.504 | |
| *Condition:Listening/Vocabulary* | 48.42 | 1, 205126 | 0.000 | *** |
| *ROI:Listening/Vocabulary* | 3.55 | 8, 205126 | 0.000 | *** |
| ***Condition:ROI:Speaking/Grammar*** | 1.95 | 8, 205126 | 0.049 | * |
| ***Condition:ROI:Listening/Grammar*** | 11.82 | 8, 205126 | 0.000 | *** |
| ***Condition:ROI:Listening/Vocabulary*** | 3.27 | 8, 205126 | 0.001 | ** |

Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs except for posterior left. Significance of this contrast across ROIs was corrected for 9 comparisons using a Bonferroni adjustment. While the time course of the response to phrase structure violations shown in **Figure 3.1** shows only a weak response during this time frame (300-500 ms), the violation effect was consistently positive – even if only weakly – throughout that range. Moreover, the significance of this positivity at all except for the posterior left ROI was in accordance with the

topography of this effect shown in **Figure 3.2**, as this represents the scalp region at which the peak of a (weak) negative response was shown. Overall, evidence for these effects in the grand averaged time course and topography of the response were weak. Nonetheless, the effect was statistically significant.

During the 300-500 ms time window, participants with lower Listening/Vocabulary scores showed a stronger positive violation effect at anterior and central midline ROIs, and those with lower Listening/Grammar scores showed a similar positive violation effect at the central midline ROI. However, the influence of the ID measures at these regions, as well as amplitude of the response, were relatively weak. Conversely, participants with higher Speaking/Grammar scores showed a stronger positive violation effect than did those with lower scores. The influence of Speaking/Grammar on response amplitude was significant at anterior left and midline, central midline, and posterior left ROIs. These effects are outlined in **Figure 4.8**.

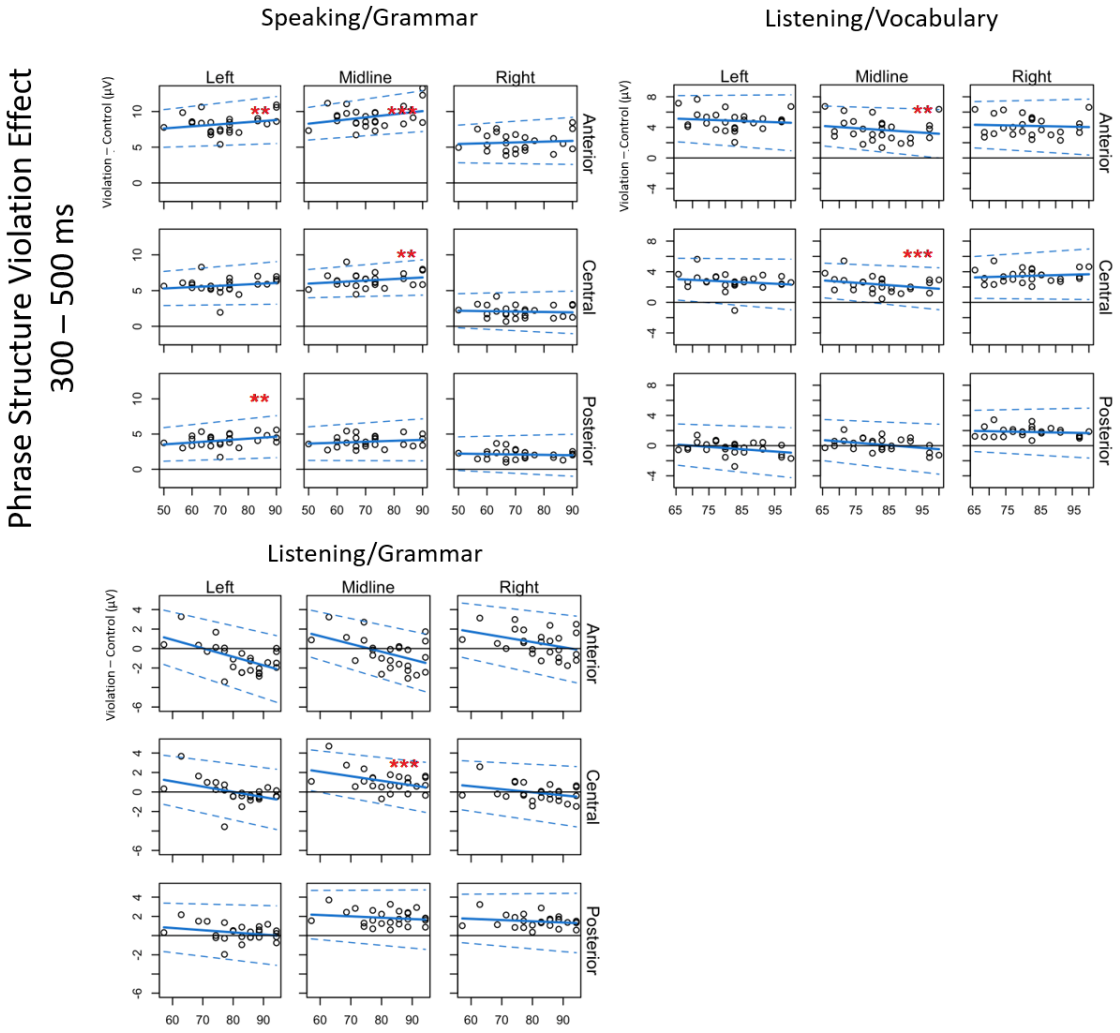


***Figure 4.8*** *Phrase structure violation effects across each ROI* in the 300-500 ms time window. *The 95% confidence intervals are shown, indicating violation effects, with a significant slope of the predictor shown (* p < .05, ** p < .01, *** p < .001) where the CI of that effect exceeds zero.*

***Table 4.7*** *Post-hoc comparison of condition (violation vs. well-formed) contrast at each ROI, and three-way interactions (modulation of the condition contrast slope by an ID measure at each ROI) in the 300-500 ms time window for phrase structure violations. Rows show any ROI where the 95% CI of the condition contrasts exceeded zero, alongside significance of the interaction at that region.*

| Interaction | ROI | Corrected Significance | Model Estimate |
|---|---|---|---|
| Condition x ROI (DOF = 205126) | Anterior Left | t = 6.63, $p < .001$, *** | 6.14 |
| | Anterior Midline | t = 7.59, $p < .001$, *** | 6.09 |
| | Anterior Right | t = 5.31, $p < .001$, *** | 4.91 |
| | Central Left | t = 5.17, $p < .001$, *** | 4.33 |
| | Central Midline | t = 7.04, $p < .001$, *** | 4.89 |
| | Central Right | t = 2.94, $p = 0.029$, * | 2.47 |
| | Posterior Left | t = 2.57, $p = 0.091$ | 2.16 |
| | Posterior Midline | t = 3.55, $p = 0.003$, ** | 2.98 |
| | Posterior Right | t = 2.99, $p = 0.025$, * | 2.51 |
| Condition x ROI x Listening/Grammar (DOF = 205326) | Central Midline | t = -7.18, $p < .001$, *** | -0.174 |
| Condition x ROI x Speaking/Grammar (DOF = 205326) | Anterior Left | t = 3.52, $p = .004$, ** | 0.131 |
| | Anterior Midline | t = 6.06, $p < .001$, *** | 0.195 |
| | Anterior Right | t = 1.27, $p = 1.00$ | 0.047 |
| | Central Left | t = 2.55, $p = .097$ | 0.086 |
| | Central Midline | t = 3.43, $p = .005$, ** | 0.096 |
| | Posterior Left | t = 3.62, $p = .003$, ** | 0.122 |
| | Posterior Midline | t = 1.75, $p = .708$ | 0.059 |
| Condition x ROI x Listening/Vocabulary (DOF = 205326) | Anterior Left | t = -1.64, $p = .894$ | -0.088 |
| | Anterior Midline | t = -3.60, $p = .003$, ** | -0.167 |
| | Anterior Right | t = -0.93, $p = 1.00$ | -0.050 |
| | Central Left | t = -2.36, $p = .165$ | -0.114 |
| | Central Midline | t = -4.44, $p < .001$, *** | -0.178 |
| | Central Right | t = 1.42, $p = 1.00$ | 0.068 |

### 4.3.7. Phrase Structure Violations, 600-800 ms

The result of AIC-driven model for phrase structure violations in 600-800 ms is shown in **Table 4.8**. Post-hoc testing of those terms in bold is shown in **Table 4.9**.

***Table 4.8*** *Model terms for responses 600-800 ms following the onset of phrase structure violations. Significance of a term is denoted using * ($p < .05$), ** ($p < .01$), or *** ($p < .001$).*

| | ***F*** | ***DOF*** | ***p*** | |
|---|---|---|---|---|
| *Condition* | 1301.49 | 1, 205110 | 0 | *** |
| *ROI* | 4.11 | 8, 205110 | 0.0001 | *** |
| *OSpan* | 0.39 | 1, 205110 | 0.5305 | |
| *Speaking/Grammar* | 7.15 | 1, 205110 | 0.0075 | ** |
| *Listening/Grammar* | 0.11 | 1, 205110 | 0.7425 | |
| *Listening/Vocabulary* | 0.23 | 1, 205110 | 0.6322 | |
| ***Condition:ROI*** | 64.70 | 8, 205110 | 0 | *** |
| *Condition:OSpan* | 45.96 | 1, 205110 | 0 | *** |
| *ROI:OSpan* | 0.82 | 8, 205110 | 0.585 | |
| *Condition:Speaking/Grammar* | 1.84 | 1, 205110 | 0.1747 | |
| *ROI:Speaking/Grammar* | 0.67 | 8, 205110 | 0.7193 | |
| *Condition:Listening/Grammar* | 956.50 | 1, 205110 | 0 | *** |
| *ROI:Listening/Grammar* | 0.56 | 8, 205110 | 0.8143 | |
| *Condition:Listening/Vocabulary* | 135.09 | 1, 205110 | 0 | *** |
| *ROI:Listening/Vocabulary* | 1.43 | 8, 205110 | 0.1772 | |
| ***Condition:ROI:OSpan*** | 6.04 | 8, 205110 | 0 | *** |
| ***Condition:ROI:Speaking/Grammar*** | 9.01 | 8, 205110 | 0 | *** |
| ***Condition:ROI:Listening/Grammar*** | 8.63 | 8, 205110 | 0 | *** |
| ***Condition:ROI:Listening/Vocabulary*** | 7.97 | 8, 205110 | 0 | *** |

Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs. Significance in this contrast was corrected for 9 comparisons using a Bonferroni adjustment. Phrase structure violations, in comparison with well-formed sentences, elicited considerable positivity across participants that was maximal at the central midline ROI, as outlined in **Table 4.9**. The timing (600-800 ms) and distribution of this effect were consistent with that of the expected P600, which phrase

structure violations were hypothesized to elicit. Three-way interactions between condition, ROI, and each of Speaking/Grammar and OSpan suggested that the amplitude of the P600 was strongest in higher-scoring participants, when compared with those with lower scores. The dependence of P600 amplitude on Speaking/Grammar score was found to be significant at all anterior ROIs as well as the central midline ROI. Similarly, the dependence of P600 amplitude on OSpan score was significant at all anterior and central ROIs.

Three-way interactions between condition, ROI, and each of Listening/Grammar and Listening/Vocabulary similarly suggested that these ID measures influenced the P600 amplitude. However, the effect of proficiency appeared to be reversed in these measures, as P600 amplitude was found instead to be largest in participants with lower Listening/Grammar and Listening/Vocabulary scores. The influence of Listening/Grammar score on P600 amplitude was found to be significant at all posterior ROIs, as well as the central midline, anterior left, and anterior midline ROIs. While the slopes of the anterior left and midline ROIs were significant, the confidence interval of the violation effect at these regions included zero at all but the most extreme Listening/Grammar scores, and this effect may not be replicable. Similarly, the influence of Listening/Vocabulary score on P600 amplitude was found to be significant at all midline ROIs and at the left posterior ROI. These effects are outlined in **Figure 4.9**.

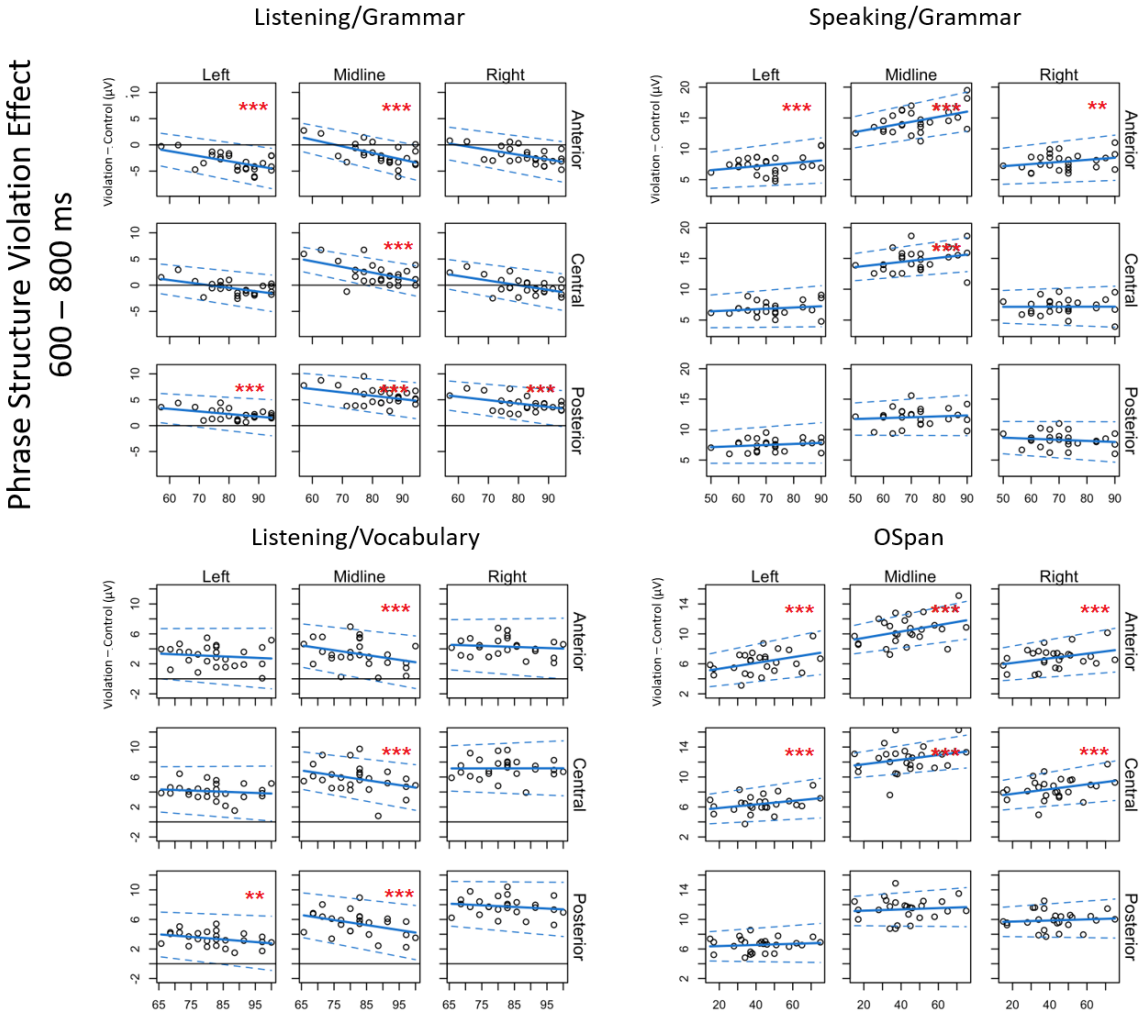


***Figure 4.9*** *Phrase structure violation effects across each ROI* in the 600-800 ms time window. *The 95% confidence intervals are shown, indicating violation effects, with a significant slope of the predictor shown (* p < .05, ** p < .01, *** p < .001) where the CI of that effect exceeds zero.*

***Table 4.9*** *Phrase structure contrast by ROI for each ID measure (600-800 ms). Rows depict contrasts where the 95% CI of the difference exceeds zero, with the significance of the interaction at this region.*

| Interaction | ROI | Corrected Significance | Model Estimate |
|---|---|---|---|
| Condition x ROI (DOF = 205110) | Anterior Left | t = 4.45, *p* < .001, *** | 4.57 |
| | Anterior Midline | t = 9.67, *p* < .001, *** | 8.60 |
| | Anterior Right | t = 5.35, *p* < .001, *** | 5.49 |
| | Central Left | t = 5.8, *p* < .001, *** | 5.38 |
| | Central Midline | t = 14.39, *p* < .001, *** | 11.08 |
| | Central Right | t = 7.65, *p* < .001, *** | 7.11 |
| | Posterior Left | t = 6.72, *p* < .001, *** | 6.24 |
| | Posterior Midline | t = 11.85, *p* < .001, *** | 11.01 |
| | Posterior Right | t = 10.29, *p* < .001, *** | 9.56 |
| Condition x ROI x Listening/Grammar (DOF = 205326) | Anterior Left | t = -9.70, p < .001, *** | -0.178 |
| | Central Midline | t = -14.54, p < .001, *** | -0.202 |
| | Posterior Left | t = -5.58, p < .001, *** | -0.093 |
| | Posterior Midline | t = -7.41, p < .001, *** | -0.122 |
| | Posterior Right | t = -7.43, p < .001, *** | -0.123 |
| Condition x ROI x Listening/Vocabulary (DOF = 205318) | Anterior Midline | t = -6.99, p < .001, *** | -0.123 |
| | Anterior Right | t = -1.33, p = 1.00 | -0.029 |
| | Central Left | t = -1.69, p = .808 | -0.032 |
| | Central Midline | t = -8.31, p < .001, *** | -0.131 |
| | Central Right | t = 0.05, p = 1.00 | -0.001 |
| | Posterior Left | t = -3.73, p = .002, ** | -0.070 |
| | Posterior Midline | t = -7.24, p < .001, *** | -0.134 |
| | Posterior Right | t = -2.35, p = .166 | -0.044 |
| Condition x ROI x OSpan (DOF = 205318) | Anterior Left | t = 6.20, p < .001, *** | 0.122 |
| | Anterior Midline | t = 7.87, p < .001, *** | 0.134 |
| | Anterior Right | t = 4.94, p < .001, *** | 0.098 |
| | Central Left | t = 4.24, p < .001, *** | 0.075 |
| | Central Midline | t = 6.63, p < .001, *** | 0.097 |
| | Central Right | t = 5.65, p < .001, *** | 0.100 |
| | Posterior Left | t = 1.41, p = 1.00 | 0.024 |
| | Posterior Midline | t = 1.64, p = .894 | 0.027 |
| | Posterior Right | t = 1.39, p = 1.00 | 0.023 |
| Condition x ROI x Speaking/Grammar (DOF = 205318) | Anterior Left | t = 4.11, p < .001, *** | 0.058 |
| | Anterior Midline | t = 10.13, p < .001, *** | 0.122 |
| | Anterior Right | t = 3.54, p = .003, ** | 0.050 |
| | Central Left | t = 2.37, p = .156 | 0.030 |
| | Central Midline | t = 7.21, p < .001, *** | 0.074 |
| | Central Right | t = 0.05, p = 1.00 | 0.001 |
| | Posterior Left | t = 2.04, p = .371 | 0.026 |
| | Posterior Midline | t = 1.74, p = .736 | 0.021 |
| | Posterior Right | t = -2.09, p = .324 | -0.026 |

## 4.4. Discussion

### 4.4.1. Overview of Objectives

The present study expanded on a framework which has been established in recent studies to characterize the link between ID measures and ERP component characteristics, with a specific focus on language-related components, the N400 and P600. There were two overarching goals of this analysis. First, we aimed to use currently-accepted best-practices to establish a baseline analysis against which investigations in subsequent chapters could be compared. Second, we leveraged developments in analytical procedures which have not been commonly used in this field to evaluate their ability to improve model fit, likelihood, and sensitivity to effects which have proven elusive or inconsistent across studies.

Prior to investigations of model improvement, we were required to determine the suitability of the commonly-used GLM for this type of experimental paradigm. While it was not expected that GLM-based approaches would be overtly inappropriate, characterizing the degree to which ERP data adhered to the required assumptions laid the groundwork for further investigation and improvements. Beyond this we aimed to explore varying degrees of random effect inclusion, up to and including theoretically-ideal random effect structures (Barr et al., 2013), while being mindful of concerns associated with specifying overly-complex random effect structures (D Bates et al., 2015). These steps allowed us to determine the most suited structure and the practical benefits of its use. Lastly, we aimed to optimize model-building procedures and use any methodological improvements to expand on our knowledge of the role of individual differences in language processing.

Individual differences were indexed in terms of vocabulary, grammatical ability, and working memory capacity. Properly modeling the effect of IDs through any technique has

important implications for several domains. These include investigating the role of proficiency in language violation processing, as well as the role of first-language proficiency in second-language processing. Effects were assessed using linear mixed effects modeling, which expands on commonly-used GLM techniques (such as ANOVA and multiple linear regression) through inclusion of random effects.

### 4.4.2. Optimizing Random Effect Structures

When determining the ideal random effect structure, we considered two sources of unpredictable variability. First, inter-participant variability was predicted, due to individual differences in factors not captured in the ID measures recorded. Each participant was expected to show variation in their mean scalp voltage for these reasons. Second, individual differences in the topography of effects were also expected, again in part due to differences in neuroanatomy but also resulting from recruitment of different neuroanatomical regions and networks during processing. In addition, individual differences in random noise (e.g., due to faulty electrodes skewing topographical distributions for specific participants) can be controlled in part through random effects. Individual differences in recruitment may occur for any number of reasons including, but not limited to, engagement in the study, system noise or participant fatigue. An exhaustive list is not possible to arrive at, and so representation of effects across our subdivision of 9 ROIs was investigated as a random effect within participants, where no underlying causes were assumed.

It is interesting to note that our analysis suggested incremental benefits across increasingly complex random effect structures, but only to a point. First, including a random by-participant intercept demonstrated a considerable improvement over a generic model including only fixed effects, as seen in the AIC, AIC weight, and model $R^2$. This is not surprising, considering

this step redefines inter-participant variability in the fixed effect as explained variance, to a degree. Second, including random ROI estimates within participants resulted in an additional improvement on all measures described above. However, this addition requires a decision: By including the random effect of ROI within the random effect of participant, the two are allowed a degree of correlation across the nested effects (i.e., a correlation of differences between participants with differences between regions). Alternatively, separating the within-participant ROI effect from the by-participant intercept forces computation of the two such that zero correlation exists between their estimates. For our purposes, both terms were categorical (i.e., the difference between any two participants had no bearing on the difference between any two others), and so this was not expected to result in any improvement. In the present data, $R^2$ was identical to four decimal places across the two methods, while the AIC (and thus AIC weight) suggested preference for allowing correlation. Indeed, doing forcing zero correlation resulted in no improvement to the model and produced a more convoluted random effect structure, which reduced the Akaike weight for this model.

#### 4.4.3. Summary of Violation Effects

For each sentence type and time window, the condition contrast (violation vs. well-formed) was significant across participants at nearly all ROIs. This condition contrast was further found to be influenced by differences between participants in scores for a number of ID measures, where in many instances a violation effect was only detected for specific ranges of scores in an ID measure. For example, semantic violations were found to elicit an N400. Critically, the amplitude of the N400 response was significantly associated with grammatical ability (Listening/Grammar), and showed stronger magnitude for lower proficiency measures, in keeping with previous research (Moreno & Kutas, 2005; Weber-Fox et al., 2003). Interestingly,

semantic violations were also associated with a degree of late posterior positivity. This finding is not uncommon, as late (600-900 ms) posterior positivity in response to semantic violations has been reported in the past, particularly in individuals with lower proficiency (Coulson & Van Petten, 2002; Juottonen et al., 1996; Kuperberg et al., 2007; Moreno & Kutas, 2005; Newman et al., 2012; Ojima, Nakata, & Kakigi, 2005; van de Meerendonk, Kolk, Chwilla, & Vissers, 2009). However, the amplitude of this effect was also found to be influenced by both Listening/Grammar and Speaking/Grammar scores, particularly in posterior regions, which to our knowledge has not been previously reported. Interestingly, a negative response was instead seen in anterior regions, and this was significantly influenced by proficiency (both Listening/Grammar and Speaking/Grammar). However, while significant, these effects were weak, and replication will help to clarify the role of these ID measures in late semantic violation effects.

While we had no specific hypotheses pertaining to responses in the 300-500 ms time window for phrase structure violations, a significant condition contrast was identified. This effect was strongest in anterior and central midline regions, but was significant in all except for the left posterior ROI. Moreover, it was found to be influenced most strongly by scores on the Speaking/Grammar and Listening/Vocabulary TOAL-3 tests. While this response was strongest in individuals with higher Speaking/Grammar scores, the reverse was true for Listening/Vocabulary scores. In both cases, ID measure influence was strongest at anterior and central midline ROIs. While significant, this conflicting pattern of proficiency effects on early responses to phrase structure violations will be important to replicate, as the interaction with Listening/Vocabulary scores in particular was related to lower overall response amplitude and weaker effect of proficiency. Considering that these are partial effects (and Listening/Grammar scores are therefore not associated with a distinctly lower-amplitude response), these findings suggest that

the interaction with Listening/Vocabulary scores is associated with less variance in the response amplitude than the interaction with Speaking/Grammar scores.

Conversely, the Speaking/Grammar interaction was associated with more variability in the response amplitude, and its influence on amplitude was larger as well. This positivity, and the positive correlation between Speaking/Grammar scores and response amplitude, dovetails with the effects of proficiency seen in the 600-800 ms time window. A P600 response was identified predominantly at central midline and posterior scalp regions, and its amplitude was strongest in individuals with higher working memory (OSpan), as reported by Nakano et al. (2010). Given that the P600 has been associated with high-level sentence repair (Nakano et al., 2010), it is unsurprising that greater working memory capacity may lend itself to this ability. Conversely, P600 amplitude in response to phrase structure violations was largest for individuals with lower (rather than higher) Listening/Vocabulary proficiency, particularly at the central midline ROI. A similar relationship was seen to a lesser degree in Listening/Grammar.

#### 4.4.4. Interpreting ID Measure Effects

It is interesting to note that in many cases, the influence of an ID measure on the amplitude of a violation effect was not seen at the central midline ROI, where the violation effects were largest, but rather ID effects were greatest in peripheral regions (i.e., more anterior or posterior regions). One interpretation of this finding is that differences in the topographical distribution (i.e., spatial extent) of a response were present when comparing low- vs. high-proficiency participants. For example, while an N400 was detected in response to semantic violations at the central midline ROI, Listening/Grammar scores did not affect its amplitude. Instead, the effect of this ID measure on N400 amplitude was significant at numerous surrounding ROIs. This pattern suggested that an N400 response was observable in the anterior

midline ROI for low-proficiency, but not high-proficiency, participants. Therefore, the effect of ID measures at regions peripheral to the peak response amplitude may represent changes in topography rather than unique regions of influence. Indeed, results suggested qualitatively different response topography for participants with lower Listening/Grammar, with the N400 response reaching farther anterior regions, while higher-scoring participants instead showed stronger positivity in this region. These results suggested that the influence of ID measures may not only be seen through strengthening of the N400 response, but also changes in the spatial extent of the response, which appear as a steeper slope in three-way interactions.

The results of our investigations largely confirmed previous reports of language proficiency having a significant effect on the latency and amplitude of the N400 and P600 (Moreno & Kutas, 2005; Nakano et al., 2010; Newman et al., 2012; Pakulak & Neville, 2010; Tanner et al., 2014; Tanner & Van Hell, 2014; Weber-Fox et al., 2003). It is important to note that the distribution of responses across Listening/Vocabulary and Listening/Grammar scores appears to fit a linear estimate reasonably well. However, some three-way interactions such as that between Listening/Grammar, condition and ROI in response to semantic violations (300-500 ms) demonstrated residuals that systematically skewed above or below the estimate across portions of the ID measure spectrum (**Figure 4.5**). This suggests a potential nonlinearity in the interaction, which cannot be captured using the present approach. In addition, considering this same interaction, participants with higher Listening/Grammar scores showed a wider range of response amplitudes than did lower-scoring participants in the leftmost anterior, central and posterior ROIs. While we discussed the potential for heteroscedasticity in these data, this type of residual variability can be the result. In these cases, a more accurate estimate may be obtained using nonparametric approaches which are more suited to handling of inconsistencies in

residuals, or nonlinear functions may be required to better fit these data. Future chapters will address each of the above concerns.

#### 4.4.5. Criteria for Determining Significance

Determining significance in three-way interactions between condition, ROI and an ID measure was based on two criteria. First, the influence of the ID measure (i.e., the slope) was required to be statistically significant in terms of the model estimate, once corrected for having performed comparisons across all ROIs (in our case, we used a Bonferroni adjustment for 9 comparisons). This was intended to ensure that only ID measures which had a considerable effect on response amplitude were deemed meaningful. However, it was still possible for statistically significant ID measure slopes to center on zero, associating this influence with a weak response amplitude. Therefore, we instituted a second requirement, whereby the 95% confidence interval of the violation effect for a significant interaction was required not to include zero at some point along the ID measure spectrum. This criterion was intended to ensure that, while the influence of an ID measure was statistically significant, the effect size of the condition contrast was also meaningful (i.e., that response amplitude was reasonably strong).

The result of these specifications was that a number of three-way interactions were deemed significant and meaningful, even in cases where the 95% CI of the violation effect only marginally exceeded zero. Furthermore, this marginal effect size was frequently only the case in a small number of participants, or even a single (highest- or lowest-scoring) participant. When combined with a statistically significant slope (albeit one which centers on zero μV), the result can suggest that the influence of an ID measure is significant, even when an effect is barely evident. This pattern was evident, for example, in the effect of Listening/Grammar on the

response to semantic violations during the 300-500 ms time window, specifically in the right midline ROI.

Qualitatively it is not difficult to see that these effects are weak, or are only visible in a minority of participants, and thus they may not replicate in a larger sample. We were therefore not concerned that the reader would misinterpret these effects to be unduly strong or meaningful. However, it does represent an intuitive inconsistency in identifying significant three-way interactions for this type of analysis. Therefore, more stringent criteria might be set for determining significant effects of this type of conditional contrast in future studies. For example, setting a required effect size that is commensurate with some minimum response amplitude of interest (rather than zero μV, as in the present study) might result in more meaningful post-hoc testing of conditional contrasts in each ROI.

#### 4.4.6. Conclusions

The present study aimed to use currently-established best practices in linear modeling procedures to develop a baseline analysis of the influence of several ID measures on N400 and P600 amplitude. This analysis was intended to provide a frame of reference for investigations using less common techniques in subsequent chapters, but also to explore two aspects of the modeling procedure: First, the impact of varied random effect structures on metrics of model quality, and second, approaches to avoiding multicollinearity. While random effects are rarely included in modeling procedures in this field, all evidence suggested that they can provide considerable gains to model quality. Moreover, while no approach to reconciling multicollinearity is free of limitations, we have presented a framework which can help to reveal those predictors which contribute most strongly to a model.

Three-way interactions between condition, ROI and individual ID measures suggested that there were inconsistencies in residual variance across scores on several ID measures. This residual distribution may have been related to nonlinearities underlying the relationship between ID measure scores and response amplitude, calling into question the ability for linear modeling techniques to adequately describe these data. However, this may be indicative of more than an inaccurate model. The two findings may represent the same problem: that it may not be possible to ideally model the present data using linear estimates, and that variance in violation effects may be most apparent at more specific ID measure ranges. Further investigation will therefore be required in more flexible approaches that are capable of describing nonlinear interactions. In the following chapter, we will discuss a robust nonlinear modeling approach to identify potential gains that it may provide.

# Chapter 5: Modeling nonlinear effects of individual differences in ERP data using generalized additive mixed modeling

## 5.1. Introduction

In our previous analysis of the relationship between several ID measures and the amplitude of N400 and P600 responses to language violations, we determined that the underlying data may be better-represented using nonlinear modeling techniques. This was reflected in scatterplots that depicted the relationship between ID measures and response amplitude (See Chapter 3). Indeed, while these relationships have been evaluated in the past, this has only been done using linear modeling techniques, despite that the present data suggest potential improvements may be found by using nonlinear approaches instead (Liang & Chen, 2014; Moreno & Kutas, 2005; Newman et al., 2012; Pakulak & Neville, 2010; Tanner, 2013; Tanner et al., 2014; Tanner & Van Hell, 2014; Weber-Fox et al., 2003). In light of this, the present chapter aims primarily to apply a nonlinear modeling technique to a similar model selection framework which was established in Chapter 4, while also investigating the impact of a number of user-defined parameters that surround model fit algorithms and approaches.

GLM-based approaches (including linear mixed effects) rely on the assumption that changes in an independent variable correspond with a linear change in the dependent variable. To date, all known investigations of ID measures on language violation processing have used these approaches (Liang & Chen, 2014; Moreno & Kutas, 2005; Newman et al., 2012; Pakulak & Neville, 2010; Tanner, 2013; Tanner et al., 2014; Tanner & Van Hell, 2014; Weber-Fox et al., 2003). However, it is unlikely that a one-to-one correspondence exists between assessment score and scalp-recorded voltage in response to violations. For example, LME suggested that

responses to semantic violations during the 300-500 ms time window might benefit from using a nonlinear fit, in that residuals were consistently above our linear fit for individuals with higher Speaking/Grammar scores, and below for those with lower scores. This very pattern was also seen in the response amplitudes outlined in Chapter 3, where only participants with lower Speaking/Grammar scores showed a negative violation effect during this time window. This information provides compelling evidence to examine the ability of nonlinear modeling techniques to improve model fit and more accurately describe the underlying relationships.

In the case that the assumption of linearity does not hold, as the present data suggest may be the case, linear regression may be unable to appropriately characterize relationships between language proficiency and violation processing altogether in certain cases. For example, Tremblay and Newman (2015) have demonstrated that parabolic effects – in which no overall trend exists, but a recognizable nonlinear pattern is evident – can be completely undetectable using linear regression. Nonlinear approaches, however, can fit a model to this type of relationship and predict scalp voltage accordingly based on dependent variable values (in our case, ID measures).

While this type of function represents a worst-case scenario for linear regression, less-dramatic nonlinearities can still prove problematic. For example, if an exponential relationship exists in the data, then using linear terms inherently cannot be completely accurate, except at intersecting points in those functions. The result must be reduced model fit, resulting in undefined variance, hindering significance testing (recall that *F* test significance hinges on a comparison of defined with undefined variance). Moreover, fitting linear terms to exponential relationships introduces error into effect size estimation at any non-intersecting points, as discrepancies between model fit for any given set of predictor coefficients (i.e., model estimate) and observed data amount to errors in generalization of findings to a population, to the degree

that observed data accurately represent that population. The most likely result of these circumstances above is uncertainty in the reported significance of model terms or *post-hoc* contrasts, and the inability to draw well-founded conclusions.

Critically, the nature of relationships between the investigated ID measures and responses to semantic or phrase structure violations investigated in the previous chapter are not yet known—a situation which is also the case for virtually any study wishing to relate ERPs to IDs given the current state of knowledge. While the above instances represent theoretical problems for linear modeling solutions, the presence of nonlinearities in these relationships have not yet been shown, and linear solutions may therefore prove adequate. As no systematic evaluation into the efficacy of these techniques in comparison with nonlinear approaches currently exists, this represents a critical next step in optimizing modeling approaches for individual differences in language processing.

The presence of nonlinearities will be investigated in this chapter using an expansion on the GLM which has been described in previous chapters, known as *generalized linear modeling*. This approach expands on our previous analysis by allowing for both linear model terms (such as those used in our LME analysis), as well smooth fits to a set of observations in specified interactions. For this analysis, we specified that the three-way interaction between sentence condition, ROI and each ID measure should be depicted as a smooth fit in order to test for nonlinearities in the influence of each ID measure on response amplitude in each region. While these fits need not be linear, the name generalized linear model comes from the linear addition of terms. Estimated responses therefore do not necessarily follow a normal distribution, or normal sources of error, which can allow modeling of data which would be problematic for approaches such as LME.

The generalized additive mixed model (GAMM; Hastie & Tibshirani, 1990; Lin & Zhang, 1999; S. Wood, 2006; S N Wood, Goude, & Shaw, 2015; Simon N. Wood, 2011) further expands on these mechanics by allowing the linking function (i.e., the function that describes the relationship between a predictor and dependent variable) to exhibit nonlinearity, rather than by describing this relationship using a linear estimate. Additive smooth terms, each accounting for partial variance in the observed data, are constructed of splines, and the method of defining nonlinearity is determined by choice of spline type. A popular choice, *cubic regression splines*, attempts to fit the partial variance of a specified effect using a series of cubic polynomials. Polynomials are required to be continuous, and each is fit to the dependent variable over the full range of the independent variable, but with the peaks of each basis function at different *x* coordinates to result in a smooth fit. The mathematical derivation of cubic regression splines is outlined elsewhere (Wood, 2006). An example of model construction using each spline type is shown in **Figure** 5.1, which has been adapted from Baayen et al. (2017).

Alternatively, *thin plate regression* splines attempt to fit observed partial variance for a specified effect through addition of increasingly-complex terms iteratively until an adequate description of the data has been reached, or until some maximum number of functions (i.e., maximum degree of complexity) have been reached. The complexity of this term is specified by the number of basis functions. Where the first and second basis function specify the intercept and slope, respectively, a third will model a parabolic dependence (opening up or down, depending on the sign of the coefficient), and increasingly 'wiggly' fits can be achieved through additional basis functions (S. Wood, 2006). While thin-plate regression splines are considered appropriate for modeling univariate smooth terms, the applicability of both to the present data was investigated (Simon N. Wood, 2003).

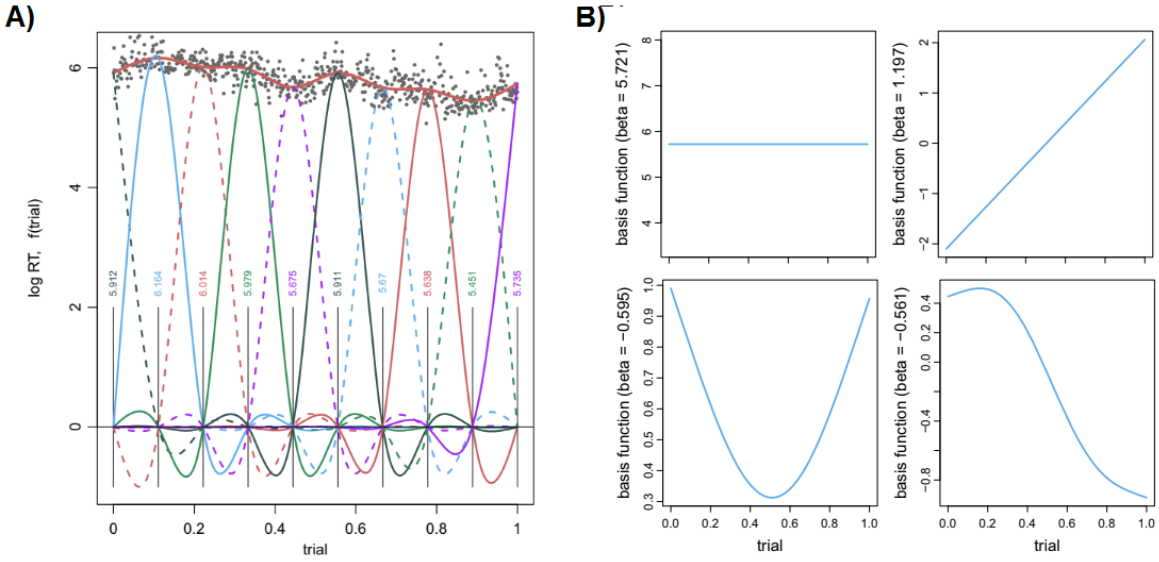


***Figure 5.1*** *Examples of model construction using each of cubic regression splines (A) or thin-plate regression splines (B) are shown, adapted from Baayen et al. (2017). Thin plate regression splines produce increasingly-complex basis functions using higher-order polynomials until their additive product adequately describes a set of observations, where the definition of adequate is governed by internal upper limitations on function complexity. Cubic regression splines instead subdivide a predictor into quantiles (joined by knots), inside which cubic polynomials are fit to fluctuations in observations.*

Regardless of the choice of spline type, the flexibility of fit to the data must be specified. Restricted cubic regression splines achieve a degree of flexibility proportionate to the number of splines desired, where *k* splines can each apply a cubic polynomial to a portion of a predictor equal 1/*k* of that predictor's range. This is specified through the number of conjoining knots, *j*, where *j* – 1 splines are permitted. This subdivision of a predictor's range into quantiles is depicted in **Figure** 5.1. When using thin plate regression splines, the weighted addition of basis functions specifies the exact shape and thus the complexity of the function. It should be noted that the complexity of a model is upwardly limited by the quantity of data. A minimum of two

points are required to produce a fit (linear or nonlinear), and more to fit a set of cubic polynomials or a series of increasingly-complex basis functions. Therefore, the density of a data set must be appropriate for the desired complexity of a model, and the two are inexorably related. In the case of either spline type, too simple a function (e.g., linear) may underrepresent the underlying relationship between two variables, and too complex (i.e., too *wiggly*) may over fit to the data and produce estimates which do not generalize. In either case the user-defined complexity of the linking function represents a trade-off between accurate description of the observed variance and generalizability to a population.

Given that GAMM requires the user to specify each of these parameters (spline type and maximum function complexity), the present chapter aimed not only to evaluate the appropriateness of nonlinear modeling techniques for relating ID measures to ERP component amplitude, but also to characterize the impact that these decisions can have on resulting fits to arrive at a robust modeling framework. Specifically, following up on evidence in previous chapters that these underlying relationships may be nonlinear, we aimed to describe whether applying nonlinear modeling techniques could provide in a better model fit when compared with LME, and provide meaningful insights into the underlying data. Importantly, the degree of nonlinearity that was expected (i.e., suggested in our previous findings) would play a role in determining the maximum complexity of models. Visual inspection of the relationships between ID measure scores and violation effect amplitude as reported in Chapter 3 suggested that while a linear estimate may be appropriate to describe the influence of several ID measures, others (e.g., Speaking/Grammar) may be best-described using asymptotic, or at most parabolic functions. Therefore, a range of complexity thresholds was used to evaluate model fit.

Following the linear modeling approach taken in the previous chapter, cortical responses were evaluated as a function of these same predictors using generalized additive mixed-effects

modeling, using the *mgcv* package (S N Wood et al., 2015) in *R version 3.4.1 (R Development Core Team, 2013)*. The previous procedure was followed as closely as possible to allow comparisons where possible, with the key difference being that models were constructed using *GAMM* instead of *LME4*. The two modeling techniques were compared in terms of model fit ($R^2$), likelihood vs. complexity using the Akaike information criterion (AIC; Akaike, 1974), and the ability to depict nuances in the violation effect over across topographical regions, latencies and ID measures. The previous chapter served as a starting point in terms of the eliminating collinear predictors, and beginning with the ideal random effect structure (Barr et al., 2013; D Bates et al., 2015). This was followed with an investigation of the most appropriate spline type, and subsequently the ideal model complexity given the volume of the present data set.

It was not obvious what impact the choice of spline type may have on model fit, and so this step was largely exploratory. However, as the maximum model complexity that a data set can support is determined primarily by its size, we expected that our present sample size of 33 participants should favor a relatively simple model. Moreover, an important consideration is that there is often no pre-existing knowledge of what level of complexity should be required to model relationships such as these. While our previous findings have suggested that it may be important to model nonlinearities, empirically determining the degree of complexity that is allowed can be difficult. Evaluating the trade-off between complexity and model parsimony can be made more challenging by the fact that models with more knots or basis functions do not require more terms to adhere to individual variance, and so approaches such as using the AIC – which penalize the numbers of terms – are likely to favor more complex models, even if they produce fits that conform so strongly to individual variance as not to be generalizable at all. Therefore, while we investigated whether this was indeed the case, we also expected that a simpler model may be required to avoid over-fitting to the point of non-generalizability.

Given these expectations and our previous findings, it was hypothesized that of the range of complexities investigated, as determined through either the number of knots or basis functions (depending on spline type), one that uses only third- or fourth-order basis functions, or the same number of knots, would be required to ideally capture the variance of the present data. It was hypothesized that relaxing the assumption of linearity and optimizing model parameters through these procedures would allow for depiction of nonlinearities in the relationship between violation effect size and ID measure, which linear modeling solutions would be incapable of describing.

In addition to evaluating the appropriateness of nonlinear modeling techniques to this type of research question, we were interested in whether the ability to describe nonlinear interactions demanded an increased volume of data, beyond what would be required to produce an adequate LME model. In particular, we were concerned that the present sample size may result in over-fitting, whereby the direction of smooth terms is guided only by a small number of participants and the effects may not replicate. To this end we investigated the impact of data set volume on model outcomes through modeling simulated data which adhered to the mean and standard deviation of the original data in each sentence condition and ROI, and which included the same ID measure scores, but contained no systematic influence of ID measures. Therefore, any detected influence of ID measures and/or nonlinearities should only be detected due to chance (reflecting error).

The likelihood of this outcome was investigated in a simulated data set equal in volume to the original data, and in a data set which contained double the number of participants across the ID measure distribution. In this case, as the simulated data were designed not to include any systematic effect of ID measures, this type of error (false positive) was considered over-fitting to random variance, and because simulated data were rendered using variance in observed data,

this should suggest that the present sample size may be problematic for models built using GAMM. Furthermore, a reduction in error rate that was found to be associated with an increase in sample size was taken as evidence that additional participants might mitigate problems associated with over-fitting. It was expected increasing the sample size in our simulated data would improve statistical power, resulting in smoother estimates of violation effect across ID measures and narrower confidence intervals. The result would be more consistent estimates of violation effect size (i.e., weaker slope), as these data were designed to contain no influence of ID measures to begin with.

## 5.2. Methods

All details pertaining to data acquisition and pre-processing are as described in Chapter 2, *Data Collection and Pre-Processing*. All subsequent details were specific to the present investigation.

### 5.2.1. Nonlinear Modeling and Visualization

A limitation of the nonlinear modeling technique used here is that it does not natively estimate the significance of two contrasted smooth terms. That is, while the predicted response to either sentences containing violations or well-formed ones can be evaluated, the difference between them (the violation effect) cannot. However, given that ID measures have been shown to selectively influence either control or violation sentences alone (Newman et al., 2012), we felt that it was important to include both in the model, rather than use the conditional contrast alone as the dependent variable. Moreover, modeling the contrast rather than each condition alone would necessitate an aggregate mean across trials and a substantial loss of statistical power. However, the decision to keep the two separate resulted in is an obstacle in comparing the significance of contrasts between LME and GAMM. Here, *F* and *p* values were therefore

evaluated for each sentence type individually while LME models were reported with *F* and *p* values for the condition contrast. This will remain a consideration in any comparisons between the two techniques.

Recall that declaring significance of a sentence-type contrast at an ROI using LME required that the 95% confidence interval of the contrast term not include zero for some level of the ID measure being considered. This stipulation was mirrored as closely as possible using GAMM through calculating the 95% confidence interval of each sentence type's smooth term, and computing a confidence interval that is the summation of the two for interpretation of the linearly subtracted condition contrast. Calculation of this contrast and confidence interval is outlined in **Figure** 5.2. Using this approach, the upper limit of the 95% confidence interval for the violation or control sentence terms are $UCI_{violation}$ or $UCI_{control}$ respectively, and the lower limit is $LCI_{violation}$ or $LCI_{control}$, the upper limit of the contrasted term's confidence interval was calculated as $UCI_{contrast} = UCI_{violation} - LCI_{control}$. Similarly, the lower limit of the contrasted term's confidence interval was calculated as $LCI_{contrast} = LCI_{violation} - UCI_{control}$. The contrasted term itself was derived through a subtraction of estimates at any given *x* coordinate ($Y_{contrast} = Y_{violation} - Y_{control}$).

Computing a summation of the two confidence intervals in the manner described above provides a valid visualization of how the confidence intervals of each term correspond to one another, but it is not equivalent to the confidence interval of a contrasted term, as was visualized using LME in the previous chapter. Rather, this calculation provides an approximation, which may vary slightly with the degree of correlation between the two terms. Therefore, while direct comparisons of the significance of contrasted terms between the two techniques was not possible (especially in regard to specific ID measure values at which a violation effect is significant), this approach served to highlight confident dissimilarities between the two sentence types.

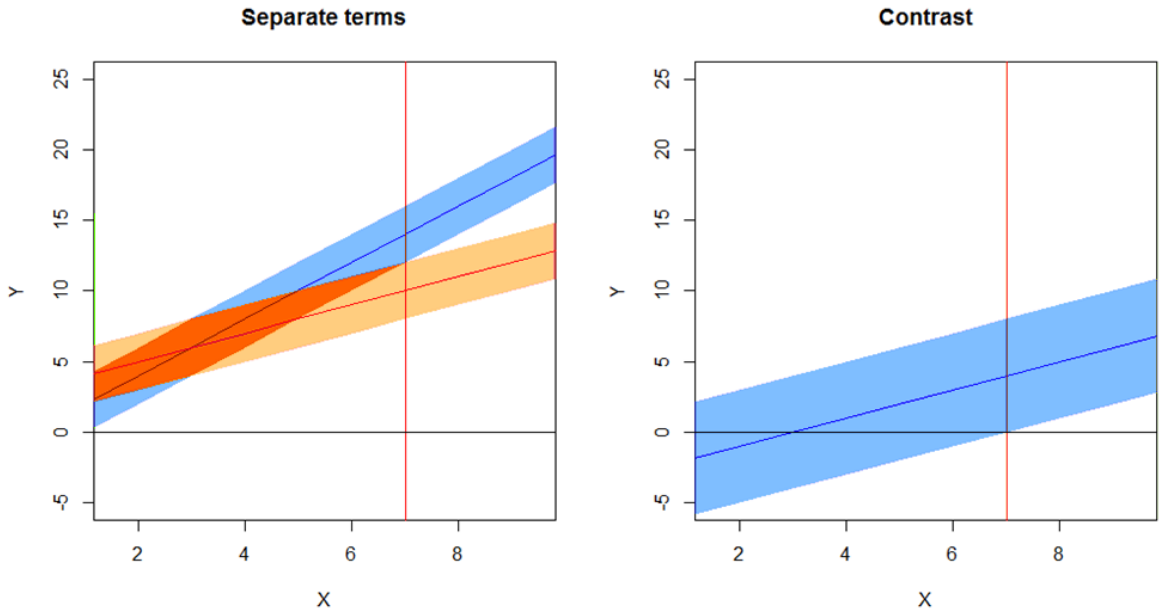


***Figure 5.2*** *On the left, two slopes are shown: Y = 2X (blue), and Y = X + 3 (orange). Each is shown with simulated 95% confidence interval of 2.0. On the right, the divergence of their confidence intervals is shown at X = 7.0 by subtracting the first function from the second, and combining their confidence intervals for any given X value. This method was used to visualize deviation of responses to violation sentences from that of well-formed sentences.*

### 5.2.2. Optimizing Model Parameters

As discussed, smooth term specification using GAMM requires that the user choose two critical parameters. The first parameter choice pertains to spline type, impacting the formulas for individual splines, which can therefore influence the shape of the function overall. Our investigations included restricted cubic regression splines and thin-plate regression splines (Wood, 2000; Wood, 2003). This step was performed as an exploratory measure, to avoid making an a priori assumption as to which was preferable in describing the data. Models fit using each of the two spline types were fit using a maximum of four knots (cubic regression splines) or four basis functions (thin-plate regression splines), and their efficacy was compared using AIC.

However, as this measure is influenced by the number of terms and the volume of data (both of which were constant between the two models), differences in AIC were directly representative of differences in the log-likelihood of the models. This type of comparison followed the logic of the same AIC-driven model selection framework outlined in Chapter 4, but here applied to the rationale of revealing the ideal mathematical derivation of functions (i.e., spline type), rather than the inclusion of any specific ID measures.

The second parameter choice concerned the highest possible degree of complexity in the linking function. In the case of cubic regression splines, this is determined through the maximum number of splines, or more specifically, through the maximum number of knots that connect splines. Recall that two knots can result in a single spline, and each additional knot can result in up to one more. In the case of thin plate regression splines, function complexity is instead specified through the maximum number of basis functions fit to the observations. Ideal function complexity is related to data set quantity/density, variability, and the nature of the predicted effect. Therefore, using whichever spline type resulted in the model with the best log-likelihood, a series of model fits were produced, each using incremental increases in complexity. Wood (2017) notes that while the choice of function complexity should not be considered critical, it should be large enough to represent the underlying truth with a reasonable degree of accuracy – and also that this criterion is subjective to the research question. Moreover, in cases where it is not obvious, Wood (2017) suggests fitting a series of models of increasing complexity. This approach was taken here. Fits were computed that used either three, four, five or six knots or basis functions. A model cannot be fit with fewer than two knots, which can produce either a linear or nonlinear spline between them. The choice of a minimum of three knots or basis functions was selected to reflect the expectation of potentially parabolic linking functions discussed previously, while still allowing for simpler (linear) functions to arise. Note that the

number of knots represents the maximum function complexity, rather than an absolute specification of complexity, and so even a specification of six knots might produce a linear fit. Our upper bound of six was selected because any fit which required more was unlikely, as no evidence suggests that increases in an ID measure scores should coincide fluctuations in ERP amplitude that would require more than six knots. As with evaluating the choice of spline type, the AIC was used to calculate the log-likelihood for each model.

For our evaluations of spline type and model complexity, responses to phrase structure violations during the 600-800 ms time window were used. These data were selected as this sentence type and time window were shown to correspond with the P600 that phrase structure violations were intended to elicit, and so their impact on fit to this response was of interest. Findings concerning which spline type was most appropriate, or the ideal complexity of the linking function, were largely identical whether describing the influence of ID measures on the P600 or N400, however, and so only the former was reported. The model building process (i.e., selection of ID measures for inclusion in significance testing) followed the same AIC-driven framework and random effect structure established in Chapter 4.

### 5.2.3. Defining Violation Effects

Once the ideal model parameters had been determined, the violation effect (violation – control condition contrast) was modeled in each of the two sentence types (semantic and phrase structure violations), for each of the time windows (300-500 ms and 600-800 ms). The procedure that was used for generation of linear models was closely followed in GAMM, with the notable exception of allowing for nonlinearities in model fit. Otherwise, however, the data used for each of the four models, and the model building process, was identical to that described in the previous chapter. Regarding addressing the collinearity of predictors, the same

pairs of predictors that were regarded as collinear in the previous chapter were tested using the same AIC selection framework here as well, and just as before only one predictor from each collinear pair was included in the model for significance testing. However, the predictors that were used in the final models (i.e., those that were not eliminated) differed because their nonlinear influence on response amplitude impacted the models' AIC values. Therefore, any predictor which had been eliminated when using LME, but which could provide a better log-likelihood when fit using a nonlinear function, might still be found significant using GAMM.

The significance of the violation effect as each ROI was evaluated using the two-way condition by ROI interaction, with the significance of post-hoc analyses corrected for 9 ROIs using the Bonferroni adjustment. Following this, the three way interactions between condition, ROI, and any ID measure (provided that interactions were significant) were similarly evaluated in each ROI, and corrected for 9 ROI comparisons. The contrast between the violation and control smooth terms was plotted in each ROI, alongside the summation of the 95% confidence intervals for these two terms, to depict where the confidence intervals of the two diverge from one another (i.e., where their summed confidence interval does not include zero μV).

Following a description of the violation effects in each time window and sentence type, we compared the final four models that were arrived at through our AIC-driven model selection process to those which were created using LME in Chapter 4. The two techniques were compared in terms of model fit ($R^2$), identifying the ability of either technique to conform to individual variance, but also using the AIC, to identify the trade-off between model parsimony and log-likelihood in each case.

### 5.2.4. Data Simulation

In order to evaluate the propensity for GAMM to over-fit to random fluctuations in individual-level variance, resulting significant ID measure influences or nonlinear fits where neither should exist, we assessed model fit in simulated data which was designed to contain no influence of ID measures, but maintained the condition contrast specific to each ROI that was found in our original data. To achieve this, we replaced the recorded scalp voltages in our original data set with a random normal distribution of responses which adhered to the mean and standard deviation of responses for each condition in each ROI. Data were generated indiscriminately of participants' ID measure scores, meaning that the resulting data set contained the ID measure scores which were included in our original data, but that any influence of ID measures on response amplitude should be due solely to random chance. In order to mirror our investigations of spline type and function complexity, simulated data were based on the distribution of responses to phrase structure violations during the 600-800 ms time window. Ten unique simulated data sets were produced, and a model corresponding to the ideal model structure for this sentence type and time window (as derived through our AIC-driven framework established in Chapter 4) was fit to each randomized data set.

Following this, an identical procedure was followed in producing a second set of ten simulated data sets, with the only difference being that this second set contained twice the number of participants. These additional participants were similarly assigned randomized responses that adhered to the mean and standard deviation of responses seen for each condition and ROI. The model structure identified above was similarly applied to these simulated data sets in order to investigate what impact a doubling of our present sample size might have on false positive identification of ID measure influence, or in over-fitting to random variance. Where these simulated data were designed to contain no influence of ID measures, a false

positive was identified as any single ROI in which a line with a zero μV slope cannot be drawn from the lowest to the highest ID measure score while falling entirely within the 95% confidence interval of the condition contrast. In addition, nonlinearities (which similarly should not occur due to the randomized nature of these data) were defined as any single ROI in which a parabolic influence of an ID measure was found. The mean number of ROIs showing either of these qualities across simulations was calculated, as well as a 95% confidence interval surrounding this estimate, to identify whether the chance in sample size resulted in a change in error rates.

## 5.3. Results

### 5.3.1. Cubic Regression vs. Thin-Plate Splines

The choice of spline type (cubic vs. thin-plate regression splines) was investigated in phrase structure violations during the 600-800 ms time window. Fits of various ID measures to the violation effect for this sentence type and time window represented a range of function shapes, showing varying degrees of nonlinearity, making them ideal for an investigation of what impact spline type might have across different types of model fits. The final model, which was arrived at through the AIC-driven model selection framework established in Chapter 4, was:

*Voltage ~ Condition:ROI:**OSpan** + Condition:ROI:**Speaking/Grammar** +*
*Condition:ROI:**Listening/Grammar** + Condition:ROI:**Listening/Vocabulary** +*
*Condition:ROI + Condition + ROI + (1 + ROI | Participant)*

This structure was applied to two models, each identical with the exception that one was fit to observations using four cubic regression splines, and the other using four basis functions to produce thin-plate regression splines. The fit of each model is depicted in **Figure 5.3**, showing the effect of spline choice on model fit for two interactions: Listening/Grammar and OSpan. While Speaking/Grammar and Listening/Vocabulary were also associated with

statistically significant influence on violation effect amplitude, they are not shown for brevity's sake. The influence of these ID measures will be discussed in detail below, following identification of model parameter outcomes.

Visually, the fit achieved by two spline types were indistinguishable. The more linear interaction of violation effect with Listening/Grammar was represented at nearly-identical magnitude and with a similar rate of change across the Listening/Grammar scores. Considering the more sinusoidal function in the interaction with OSpan scores, the two spline types similarly resulted in model fits with no substantial or apparent differences. Quantifying model fit through AIC revealed that, while differences were qualitatively negligible, cubic regression splines were preferred to thin-plate regression splines, with AIC improved by 14 (where 5 is considered significant; Akaike, 1974), and the AIC Weight associated with cubic regression splines equal to 0.999 out of a possible 1.0. Unsurprisingly, there was no difference in the significance of effects. Therefore, while no qualitative differences between the two could be discerned, the model built using restricted cubic regression splines resulted in a better log-likelihood. All subsequent models, including those used for complexity evaluations, identification of violation effects and data simulations were therefore created using cubic regression splines.

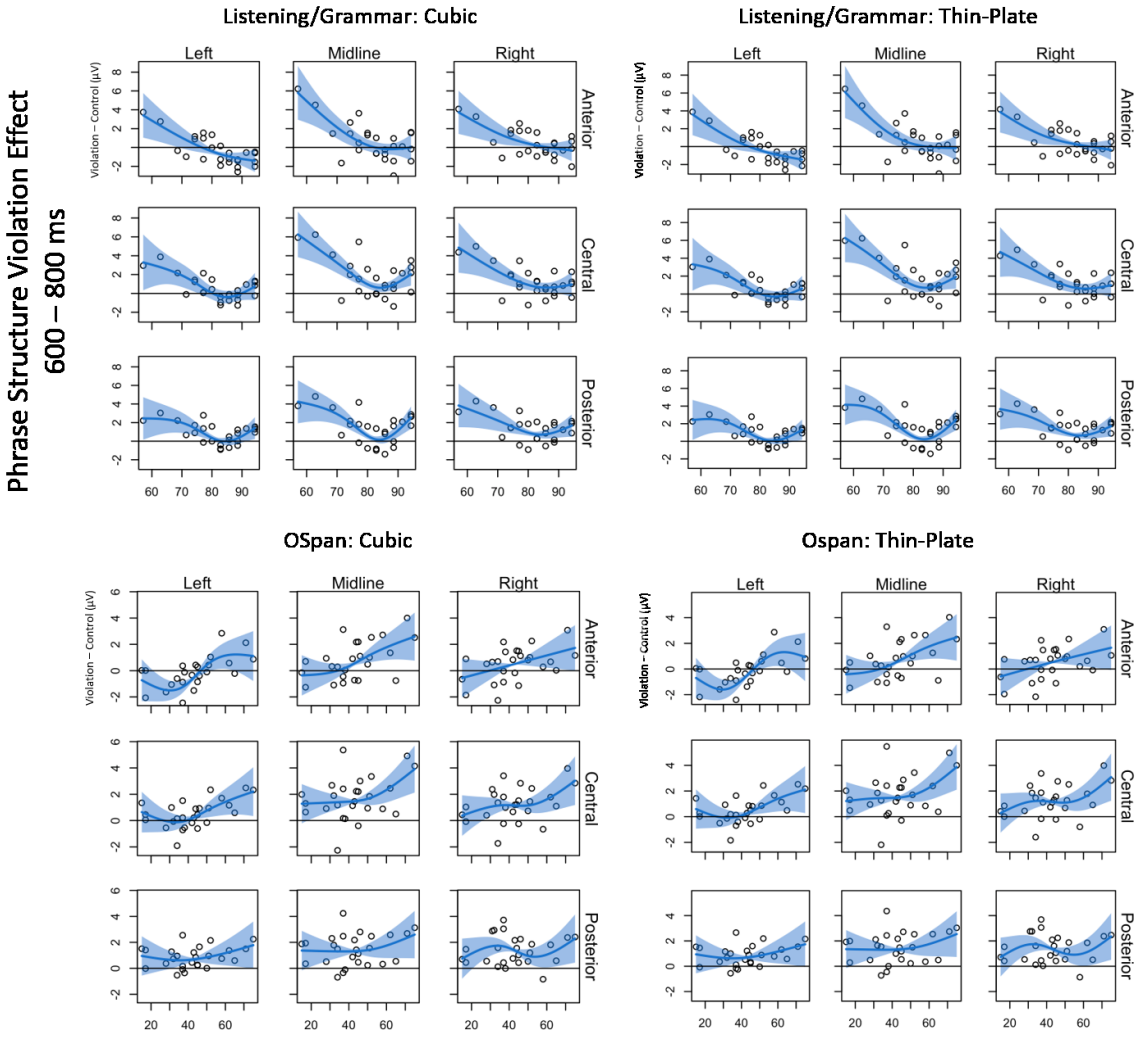


***Figure 5.3*** *The phrase structure violation effect 600-800 ms following onset of the violating word. Interactions with Listening/Grammar (top) and OSpan (bottom) are shown, each fit using cubic (left) and thin-plate (right) regression splines.*

### 5.3.2. Specifying Function Complexity

The determination of maximum function complexity, defined through the number of knots that connect them, was hindered by a lack of any specific *a priori* knowledge of function shape in the interactions of interest. That is, the nature of the dependence of the violation effect on any ID measure in each ROI could not be known beforehand. Therefore, the outcome of this

parameter was addressed through creating a series of increasingly-complex models to determine their impact on function shape, fit, and model likelihood. As with the choice of spline type, we considered phrase structure violations in the 600-800 ms time window. The final model, which was arrived at through our AIC-driven model selection process, was identical to that used for the above investigation of splice choice outcomes. However, we were primarily interested in the impact of the number of knots on the propensity to produce more sinusoidal functions than more linear ones, given that our scatterplots in Chapter 3 did not suggest such a relationship between any of our ID measures and the violation effect amplitude should be likely. We therefore focused on the fit of Speaking/Grammar scores to violation effect amplitude during this time frame, as this was the ID measure which showed the most strongly sinusoidal effect in the model across the range of complexities that were investigated

**Figure 5.4** depicts the fits of four models, ranging from three to a maximum of six knots. It is important to note that this does not necessarily reflect the number of knots that were included in the final model fit, but instead the maximum number, as determined by the *gam* function used (from the *R* package *mgcv*). The process of determining maximum function complexity is known as generalized cross-validation. This procedure attempts to evaluate the complexity of the model alongside the fit to the data in an attempt to strike a balance between simplicity and accuracy. While additional knots necessarily improve fit, this step was intended to penalize overall fit against variance in the fit of individual splines, but its performance may be variable in smaller data sets (Wood, 2006; Wood, 2008; Wood, 2011). This function is conceptually comparable to the AIC's penalization of model likelihood against the number of model terms, and similarly aims to arrive at the most parsimonious model. Fewer than the maximum number of knots in any given model are frequently shown for this reason.

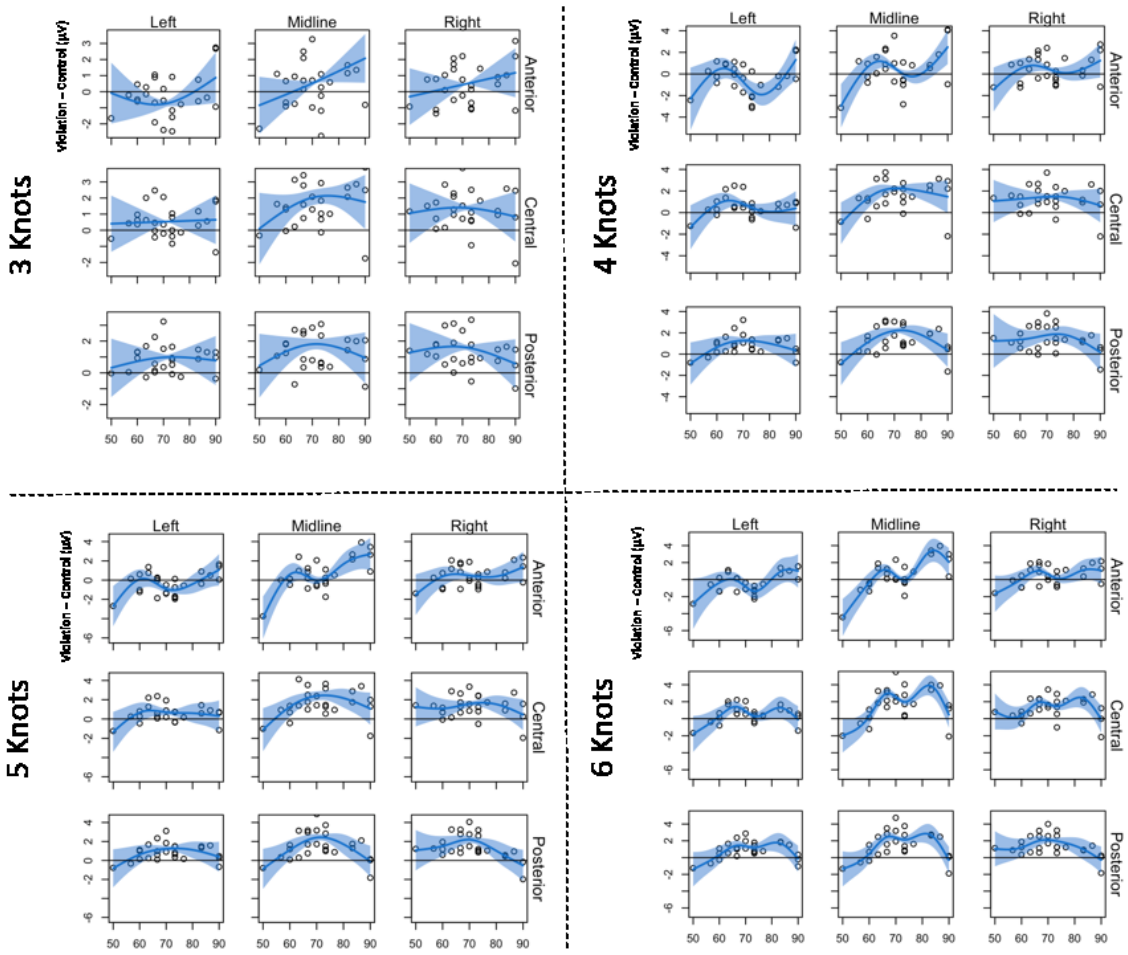


***Figure 5.4*** *Phrase structure violations in the 600-800 ms time window, with the interaction between violation effect and Speaking/Grammar shown. A separate model was created for each of three through a maximum of six potential knots.*

As the number of knots increased, fit of the model to the observations increased to the point that nearly every observation fell within the 95% confidence interval and strong fluctuations in predicted response appeared. This exemplifies that additional splines necessarily decrease residual variance and drive increases in $R^2$. Generalized cross-validation at times reduced the number of splines in an interaction, for example resulting in occasionally linear fits when three knots were permitted, or a single parabolic fit when four were permitted. However, in numerous instances the predicted responses – particularly when a maximum of five or six

knots were permitted - appeared to strongly confirm to individual data points, which may reflect over-fitting to the data. That is, we suspected that increases in Speaking/Grammar scores were likely not truly associated with repeated intermittent increases and decreases in violation effect amplitude. Moreover, our findings in Chapter 3 did not suggest this should be the case either, which appeared best-suited for fits to asymptotic or parabolic functions. This characteristic of strongly conforming to individual data points represents a considerable problem in empirically determining the appropriate number of splines in any data set. Moreover, evaluating such increasingly-complex models using the AIC (or other measures which apply penalties for the number of terms, such as BIC) is not appropriate. While each additional knot describes observed variance in increasing detail, the number of terms (i.e., smooths) is unchanged with model complexity, and no further penalty is applied during AIC calculation. This relationship is shown in ***Table 5.1***. As a result, the AIC will always be strongest in the most complex model, which is always preferred at an AIC weight of precisely 1.0 out of a maximum of 1.0.

***Table 5.1*** *Phrase structure violation effect in the 600-800 ms time window, modeled with the maximum number of knots ranging from three to eight. With additional knots, AIC improves incrementally and the AIC weight necessarily prefers the most complex model.*

| *No. Knots* | *DOF* | *AIC* | *AIC Weight* |
|---|---|---|---|
| *3* | 303.38 | 1319157 | 0 |
| *4* | 338.42 | 1318865 | 0 |
| *5* | 357.86 | 1318477 | 0 |
| *6* | 383.8 | 1318312 | 100% |

Reliance on generalized cross-validation to simplify fits was necessary as the AIC could not meaningfully distinguish between the adequacy of models of varying complexity. Critically, however, generalized cross-validation was unable to adequately simplify models in which five or six knots were permitted to produce fits which our previous findings suggested might be reasonable. Where other methods to reduce model complexity (e.g., AIC) were not appropriate, no empirical means could be established to determine the appropriate trade-off between model complexity and generalizability.

This characteristic over-fitting to observations is a known problem for *gam* (Wood, 2008), and while generalized cross-validation can reduce its impact under ideally-distributed residuals and given larger sample sizes (what sample size is sufficient is not clear), neither our residuals nor our sample size may be sufficient to rely on the procedure to arrive at the ideal function complexity. We therefore opted for simpler models, specifying a maximum of four knots, given that the relationships identified in Chapter 3 did not suggest that they should require more than three splines to be sufficiently described. This number was selected to allow for the parabolic linking function which might be required to describe these relationships, but with an additional linear (or nearly linear) spline, should any portion of an ID measure spectrum not be associated with variance in response amplitude.

#### 5.3.3. Semantic Violations, 300-500 ms

In accordance with our findings regarding spline choice and function complexity, all models describing violation effects were created using restricted cubic regression splines with a maximum of four knots in any smooth term. The optimal model (as determined by AIC) describing responses to semantic violations in the 300-500 ms time window was:

*Voltage ~ Condition:ROI:**OSpan** + Condition:ROI:**Speaking/Grammar** +*

*Condition:ROI:**Listening/Grammar** + Condition:ROI:**Listening/Vocabulary** +*

*Condition:ROI + Condition + ROI + (1 + ROI | Participant)*

ID measures that were included in three-way interactions with condition and ROI indicate those that were likely to affect violation processing. These measures are denoted in bold and were modeled as smooth terms. Effects and interactions are discussed below.

The model included a significant effect of condition ($F(1, 184 = 396.13$, $p < .001$), ROI ($F(8, 184 = 17.56$, $p < .001$), and their interaction ($F(8, 184) = 351.22$, $p < .001$). Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs, where significance was corrected for 9 comparisons using a Bonferroni adjustment. This interaction revealed a positive violation effect at all anterior ROIs, and a negative effect at all central and posterior ROIs. The central-parietal distribution of this negative effect was consistent with the expected N400. Violation effects were seen to have significant interactions with ROI and each of four ID measures: Listening/Grammar, Speaking/Grammar, Listening/Vocabulary, and OSpan. These effects are outlined in **Figure** 5.5 and **Table** 5.2.

The overall positive violation effect at anterior ROIs, and negative effect at central and posterior ROIs, was consistently depicted in three-way interactions between condition, ROI and each of the ID measures. Overall, participants with higher Listening/Grammar scores showed a stronger positive anterior response to semantic violations during this time window, as well as a higher-amplitude N400. Conversely, participants with lower Listening/Grammar scores showed weaker responses. Participants with higher Speaking/Grammar scores similarly showed a stronger positive violation effect in anterior ROIs, though Speaking/Grammar scores did not appear to strongly affect N400 amplitude at central or posterior ROIs.

Overall, the opposite pattern was seen in the three-way interaction between condition, ROI and Listening/Vocabulary, as lower-scoring participants showed both a stronger positive response in anterior regions, with no strong influence of Listening/Vocabulary scores on N400 amplitude depicted. This pattern was also noted in participants with lower OSpan scores, who showed a stronger negative response in anterior ROIs, but limited influence of OSpan on N400 amplitude.

Each of these three-way interactions to some degree showed sinusoidal fluctuations in the relationship between ID measures and the amplitude of the violation effect, whereby increases in a score on any of the ID measures was associated with intermittent fluctuations in the predicted amplitude of the violation effect. In some interactions, for example the anterior midline ROI for the interaction between condition, ROI and Speaking/Grammar scores the strong curvature of this fit appeared to be guided by small numbers of individuals. The effect of sample size on the shape of smooth term fit will be considered below. However, in some cases such as the posterior midline ROI for this same interaction, it should be noted that the 95% CI of the contrast could also contain an estimate with zero slope, suggesting that the nonlinear fit depicted might not be required to describe the interaction.

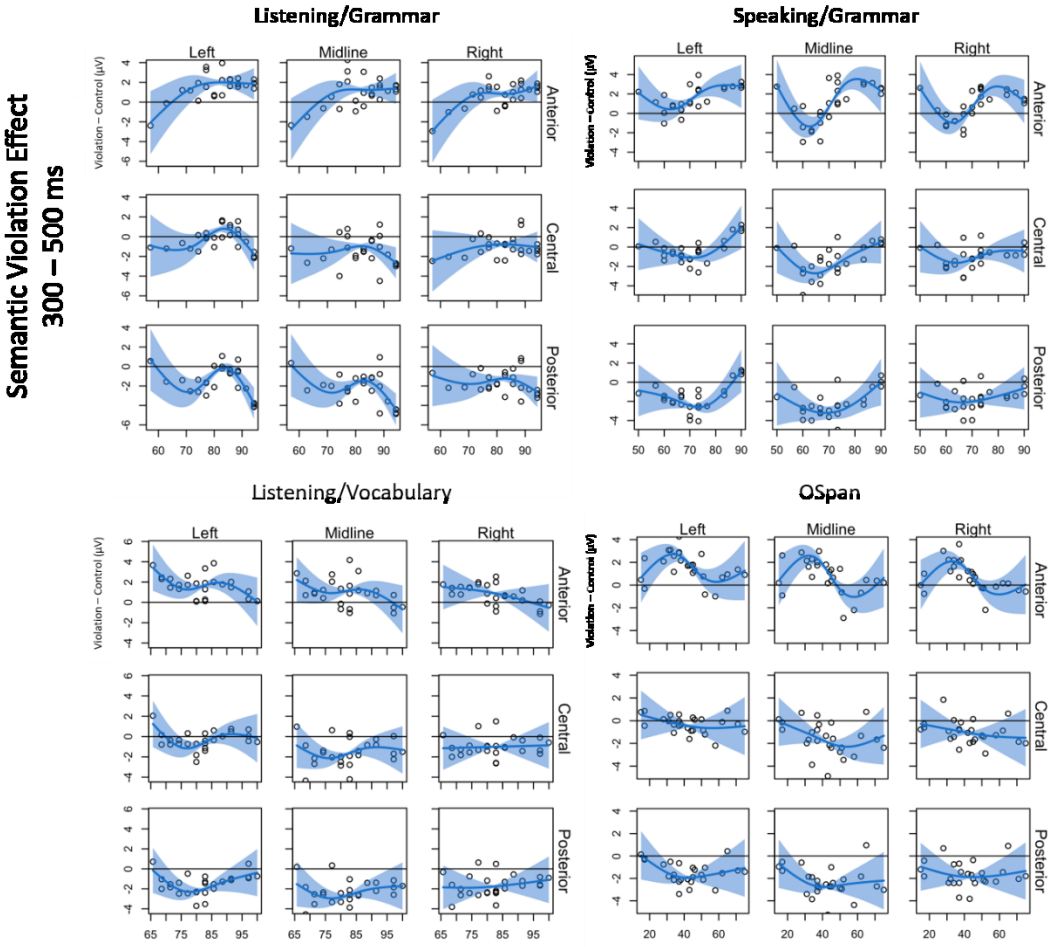


***Figure 5.5*** *Semantic violation effects in the 300-500 ms time window. The combined 95% confidence intervals of the sentence type estimates are shown, indicating where the two significantly diverge.*

***Table** 5.2 List of condition effects (violation vs. control) at each ROI across participants, and significant three-way interactions between condition, ROI, and ID measures. Effects are limited to semantic violations in the 300-500 ms time window.*

| Interaction | ROI | Violation – Control Contrast | |
|---|---|---|---|
| Condition × ROI | Anterior Left | t(201160) = 19.9, p = < .001, *** | |
| | Anterior Midline | t(201160) = 14.08, p < .001, *** | |
| | Anterior Right | t(201160) = 9.69, p < .001, *** | |
| | Central Left | t(201160) = -4.52, p < .001, *** | |
| | Central Midline | t(201160) = -25.67, p < .001, *** | |
| | Central Right | t(201160) = -13.96, p < .001, *** | |
| | Posterior Left | t(201160) = -20.92, p < .001, *** | |
| | Posterior Midline | t(201160) = -32.49, p < .001, *** | |
| | Posterior Right | t(201160) = -23.03, p < .001, *** | |
| **Interaction** | **ROI** | **Control Sentences** | **Violation Sentences** |
| Condition × ROI × Listening/Grammar | Anterior Left | F(2.7, 183.58) = 14.56, p < .001 *** | F(1, 183.58) = 0.1, p = 1.00 |
| | Anterior Midline | F(2.8, 183.58) = 19.48, p < .001 *** | F(1, 183.58) = 0.05, p = 1.00 |
| | Anterior Right | F(2.83, 183.58) = 12, p < .001 *** | F(1, 183.58) = 0.96, p = 1.00 |
| | Central Left | F(2.95, 183.58) = 23.65, p < .001 *** | F(1, 183.58) = 0.03, p = 1.00 |
| | Central Midline | F(2.86, 183.58) = 11.8, p < .001 *** | F(1, 183.58) = 0.48, p = 1.00 |
| | Posterior Left | F(2.98, 183.58) = 37.52, p < .001 *** | F(1, 183.58) = 2.71, p = 1.00 |
| | Posterior Midline | F(1, 183.58) = 0.92, p = 1.00 | F(2.96, 183.58) = 19.55, p < .001 *** |
| | Posterior Right | F(1, 183.58) = 0.54, p = 1.00 | F(2.83, 183.58) = 4.88, p = 0.034 * |
| Condition × ROI × Listening/Vocabulary | Central Left | F(1.03, 183.58) = 0.04, p = 1.00 | F(2.93, 183.58) = 16.16, p < .001 *** |
| | Central Midline | F(1, 183.58) = 0, p = 1.00 | F(2.9, 183.58) = 9.23, p < .001 *** |
| | Posterior Left | F(2.88, 183.58) = 21.71, p < .001 *** | F(1, 183.58) = 0.01, p = 1.00 |
| | Posterior Midline | F(2.82, 183.58) = 9.45, p < .001 *** | F(1, 183.58) = 0.07, p = 1.00 |
| Condition × ROI × OSpan | Anterior Left | F(2.96, 183.58) = 19, p < .001 *** | F(1, 183.58) = 0.95, p = 1.00 |
| | Anterior Midline | F(1, 183.58) = 0.5, p = 1.00 | F(2.98, 183.58) = 40.49, p < .001 *** |
| | Anterior Right | F(1, 183.58) = 0.07, p = 1.00 | F(2.96, 183.58) = 21.73, p < .001 *** |
| | Central Midline | F(2.78, 183.58) = 17.34, p < .001 *** | F(1, 183.58) = 0.5, p = 1.00 |
| | Posterior Left | F(1, 183.58) = 3.39, p = 1.00 | F(2.73, 183.58) = 14.67, p < .001 *** |
| | Posterior Midline | F(2.76, 183.58) = 13.49, p < .001 *** | F(1, 183.58) = 0.4, p = 1.00 |
| Condition × ROI × Speaking/Grammar | Anterior Left | F(1, 183.58) = 0, p = 1.00 | F(2.91, 183.58) = 11.02, p < .001 *** |
| | Anterior Midline | F(1, 183.58) = 0.19, p = 1.00 | F(2.99, 183.58) = 60.41, p < .001 *** |
| | Anterior Right | F(1, 183.58) = 0.27, p = 1.00 | F(2.98, 183.58) = 34.87, p < .001 *** |
| | Central Left | F(2.83, 183.58) = 17.69, p < .001 *** | F(1, 183.58) = 0.37, p = 1.00 |
| | Central Midline | F(2.93, 183.58) = 26.51, p < .001 *** | F(1.04, 183.58) = 1.3, p = 1.00 |
| | Posterior Left | F(2.91, 183.58) = 27.94, p < .001 *** | F(1, 183.58) = 0, p = 1.00 |
| | Posterior Midline | F(2.58, 183.58) = 5.95, p = 0.013 * | F(1.62, 183.58) = 0.48, p = 1.00 |
| | Posterior Right | F(2.57, 183.58) = 6.34, p = 0.039 * | F(1, 183.58) = 0.02, p = 1.00 |

### 5.3.4. Semantic Violations, 600-800 ms

The model describing responses to semantic violations in the 600-800 ms time window, which was arrived at through our AIC-driven term selection, was:

*Voltage ~ Condition:ROI:**OSpan** + Condition:ROI:**Speaking/Grammar** + Condition:ROI:**Listening/Grammar** + Condition:ROI:**Listening/Vocabulary** + Condition:ROI + Condition + ROI + (1 + ROI | Participant)*

The model included a significant effect of condition ($F(1, 179 = 25.77$, $p < .001$), and an interaction between condition and ROI ($F(8, 179) = 100.01$, $p < .001$). Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs except for anterior right. Significance was corrected for 9 comparisons using a Bonferroni adjustment. Overall, this condition contrast was positive at all ROIs, and showed significant negativity only in the anterior left ROI. This pattern was reflected in the three-way interactions between condition, ROI and each of the ID measures. Specifically, participants with lower Listening/Grammar scores showed a stronger positive response at central midline, posterior midline and posterior right ROIs, but no significant influence of Listening/Grammar in other ROIs. Participants with lower Speaking/Grammar scores showed a negative response at anterior left and anterior midline ROIs, but similarly no influence of Speaking/Grammar scores on response amplitude was found in other ROIs. These effects are outlined in **Figure** 5.6 and **Table** 5.3.

Listening/Vocabulary scores were not found to influence the amplitude of the violation effect at any ROI, as while nonlinearities were seen in the smooth term for each ROI, the 95% CI of this contrast either included zero μV at all points, or could contain a zero-slope line at all points, suggesting that these nonlinearities may not be required to describe the relationship.

However, higher OSpan scores were associated with a stronger amplitude positive response at all midline and posterior ROIs. This interaction also demonstrated a degree of sinusoidal curvature, conforming to a handful of the lowest- and highest- scoring participants, particularly at the posterior left and central midline ROIs.

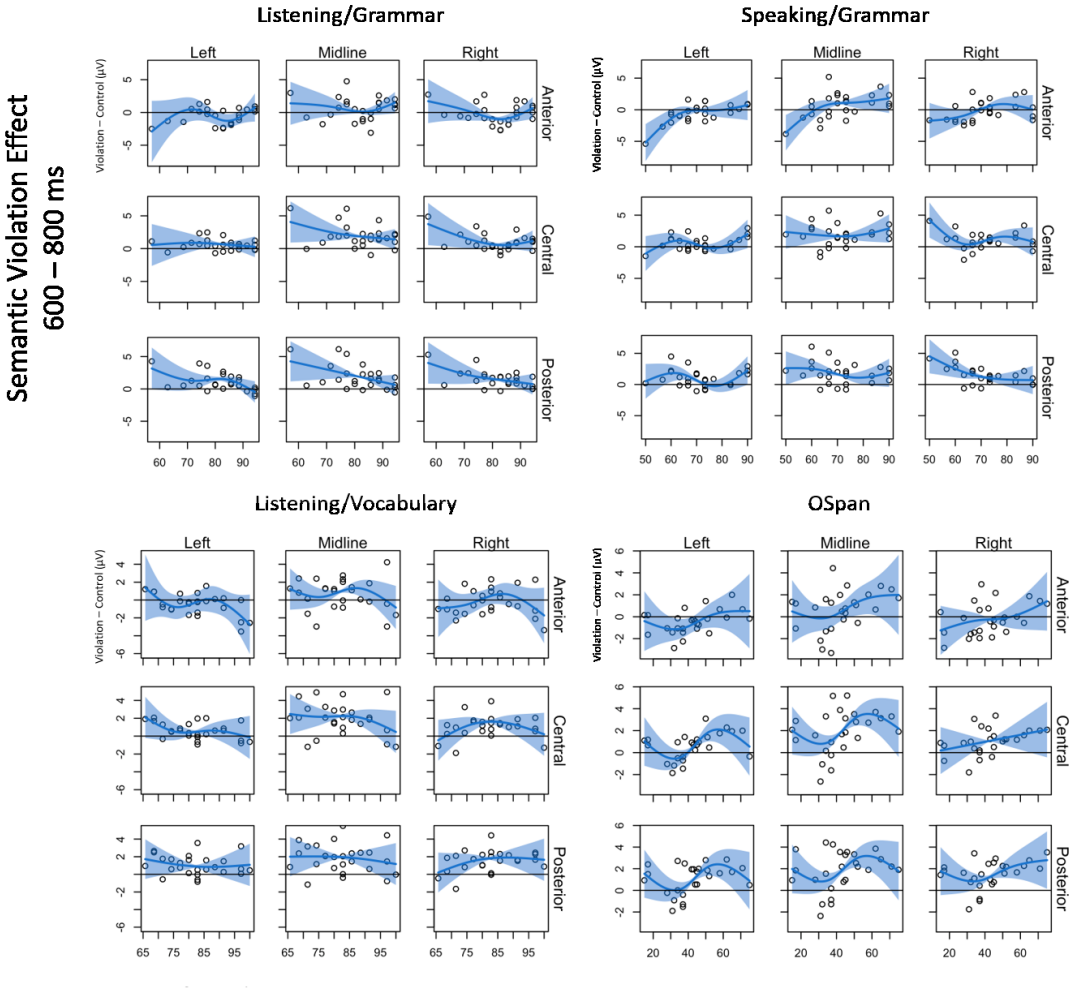


***Figure 5.6*** *Semantic violation effects in the 600-800 ms time window. The combined 95% confidence intervals of the sentence type estimates are shown, indicating where the two significantly diverge.*

***Table 5.3*** *List of condition effects (violation vs. control) at each ROI across participants, and significant three-way interactions between condition, ROI, and ID measures. Effects are limited to semantic violations in the 600-800 ms time window.*

| Interaction | ROI | Violation – Control Contrast | |
|---|---|---|---|
| Condition × ROI | Anterior Left | t(201165) = -5.08, p < .001, *** | |
| | Anterior Midline | t(201165) = 8.36, p < .001, *** | |
| | Anterior Right | t(201165) = -2.16, p = .275 | |
| | Central Left | t(201165) = 7.29, p < .001, *** | |
| | Central Midline | t(201165) = 28.57, p < .001, *** | |
| | Central Right | t(201165) = 12.67, p < .001, *** | |
| | Posterior Left | t(201165) = 12.72, p < .001, *** | |
| | Posterior Midline | t(201165) = 22, p < .001, *** | |
| | Posterior Right | t(201165) = 18.18, p < .001, *** | |
| **Interaction** | **ROI** | **Control Sentences** | **Violation Sentences** |
| Condition × ROI × Listening/Grammar | Anterior Midline | F(1, 178.82) = 5.83, p = 0.283 | F(2.77, 178.82) = 8.34, p < .001 *** |
| | Anterior Right | F(1, 178.82) = 4.1, p = 0.771 | F(2.81, 178.82) = 13.34, p < .001 *** |
| | Central Right | F(1, 178.82) = 0.85, p = 1.00 | F(2.59, 178.82) = 9.36, p < .001 *** |
| | Posterior Left | F(2.84, 178.82) = 4.91, p = 0.041 * | F(1, 178.82) = 0.86, p = 1.00 |
| Condition × ROI × Listening/Vocabulary | Anterior Midline | F(1, 178.82) = 1.55, p = 1.00 | F(2.9, 178.82) = 11.34, p < .001 *** |
| | Central Midline | F(1, 178.82) = 2.43, p = 1.00 | F(2.68, 178.82) = 7.51, p = 0.001 *** |
| | Central Right | F(1, 178.82) = 0.06, p = 1.00 | F(2.69, 178.82) = 15.65, p < .001 *** |
| Condition × ROI × OSpan | Central Left | F(1, 178.82) = 0.01, p = 1.00 | F(2.96, 178.82) = 16.67, p < .001 *** |
| | Central Midline | F(2.97, 178.82) = 22.26, p < .001 *** | F(1, 178.82) = 6.84, p = 0.16 |
| | Posterior Left | F(2.95, 178.82) = 14.1, p < .001 *** | F(1, 178.82) = 0.87, p = 1.00 |
| | Posterior Midline | F(2.94, 178.82) = 12.07, p < .001 *** | F(1.04, 178.82) = 3.98, p = 0.749 |
| | Posterior Right | F(1.01, 178.82) = 0.85, p = 1.00 | F(2.8, 178.82) = 8.46, p < .001 *** |
| Condition × ROI × Speaking/Grammar | Anterior Left | F(2.86, 178.82) = 11.75, p < .001 *** | F(1, 178.82) = 7.29, p = 0.124 |
| | Anterior Midline | F(2.83, 178.82) = 12.46, p < .001 *** | F(1, 178.82) = 10.91, p = 0.017 * |
| | Anterior Right | F(2.78, 178.82) = 4.88, p = 0.023 * | F(1, 178.82) = 2.2, p = 1.00 |
| | Central Left | F(1, 178.82) = 0.02, p = 1.00 | F(2.93, 178.82) = 12.81, p < .001 *** |
| | Central Midline | F(1, 178.82) = 2.41, p = 1.00 | F(2.49, 178.82) = 6.54, p = 0.004 ** |
| | Central Right | F(2.95, 178.82) = 19.07, p < .001 *** | F(1.02, 178.82) = 0.09, p = 1.00 |
| | Posterior Left | F(2.94, 178.82) = 15.18, p < .001 *** | F(1, 178.82) = 0.43, p = 1.00 |
| | Posterior Midline | F(2.75, 178.82) = 4.85, p = 0.026 * | F(1, 178.82) = 0.02, p = 1.00 |
| | Posterior Right | F(2.53, 178.82) = 7.43, p = 0.003 ** | F(1.02, 178.82) = 1.75, p = 1.00 |

### 5.3.5. Phrase Structure Violations, 300-500 ms

The model describing responses to phrase structure violations at 300-500 ms, which was arrived at through our AIC-driven term selection, was:

*Voltage ~ Condition:ROI:**OSpan** + Condition:ROI:**Speaking/Grammar** + Condition:ROI:**Listening/Grammar** + Condition:ROI:**Listening/Vocabulary** + Violation:ROI + Violation + ROI + (1 + ROI | Participant)*

The model included a significant effect of condition ($F(1, 181 = 6.57$, $p = .013$), ROI ($F(8, 181=4.41$, $p < .001$), and their interaction ($F(8, 181) = 21.98$, $p < .001$). Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs except for anterior left, central midline, and posterior midline, where significance was corrected for 9 comparisons using a Bonferroni adjustment. The violation effect showed significant negativity at the central left and posterior left ROIs, but was positive in all others.

The amplitude of phrase structure violation effects were seen to interact with four ID measures across condition and ROI: Listening/Grammar, Speaking/Grammar, Listening/Vocabulary, and OSpan. Specifically, participants with lower Listening/Grammar scores demonstrated a significant positive violation effect in left anterior ROI. No influence of Listening/Grammar scores on response amplitude was seen in other ROIs. Concerning each of Speaking/Grammar, Listening/Grammar and OSpan scores, none were found to significantly influence the phrase structure violation effect amplitude during this time window. While these ID measures occasionally showed divergence of the confidence intervals of the two sentence types at mid-range scores only, this was primarily due to narrow confidence intervals near the

center of the ID measures' distributions, and otherwise no overall trend was evident. These effects are outlined in **Figure** 5.7 and ***Table*** *5.4*.

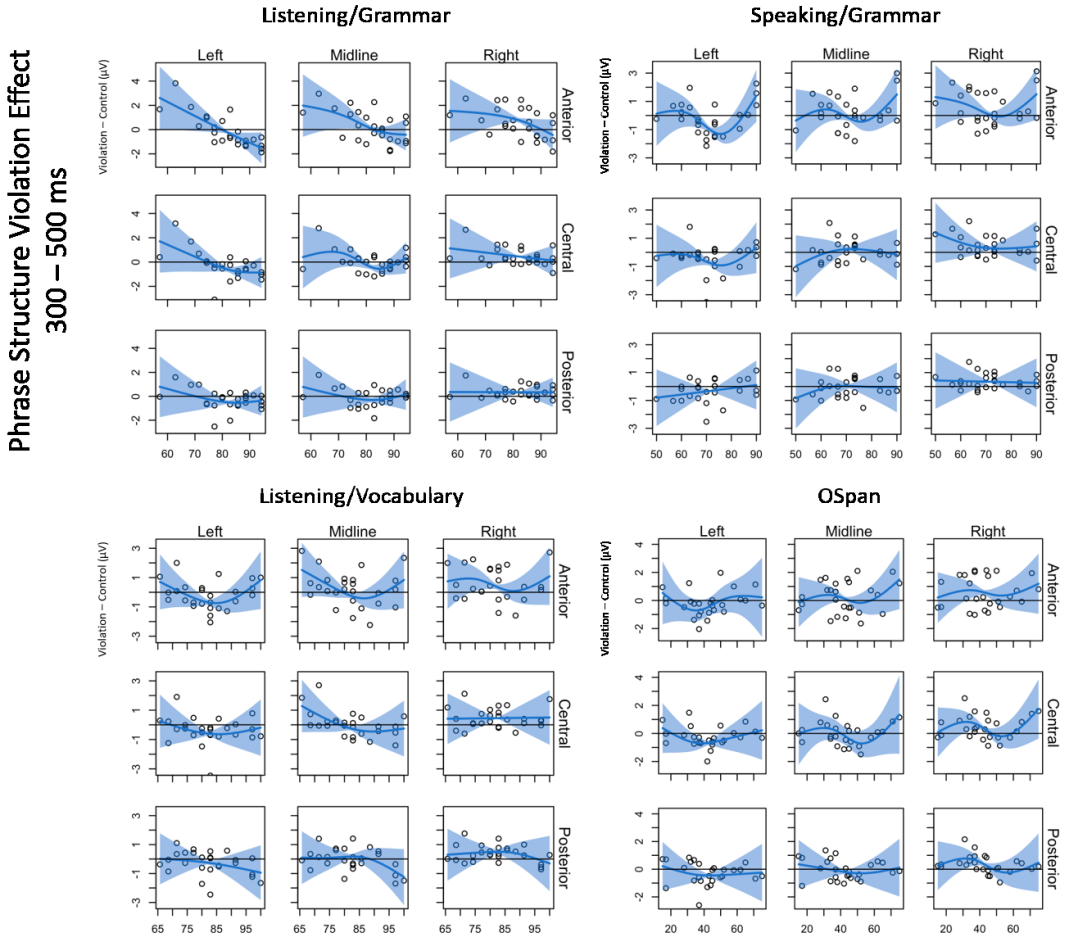


***Figure 5.7*** *Phrase structure violation effects in the 300-500 ms time window. The combined 95% confidence intervals of the sentence type estimates are shown, indicating where the two significantly diverge.*

***Table** 5.4 List of condition effects (violation vs. control) at each ROI across participants, and significant three-way interactions between condition, ROI, and ID measures. Effects are limited to phrase structure violations in the 300-500 ms time window.*

| Interaction | ROI | Violation – Control Contrast | |
|---|---|---|---|
| Condition × ROI | Anterior Left | t(205115) = -2.56, p = 0.093 | |
| | Anterior Midline | t(205115) = 3.43, p = 0.005, ** | |
| | Anterior Right | t(205115) = 7.32, p < .001, *** | |
| | Central Left | t(205115) = -5.39, p < .001, *** | |
| | Central Midline | t(205118) = 0.18, p = 1.00 | |
| | Central Right | t(205115) = 6.3, p < .001, *** | |
| | Posterior Left | t(205115) = -4.41, p < .001, *** | |
| | Posterior Midline | t(205115) = -1.06, p = 1.00 | |
| | Posterior Right | t(205115) = 4.69, p < .001, *** | |
| **Interaction** | **ROI** | **Control Sentences** | **Violation Sentences** |
| Condition × ROI × Listening/Grammar | Central Midline | F(2.9, 180.91) = 7.99, p < .001 *** | F(1, 180.91) = 0.01, p = 1.00 |
| Condition × ROI × Listening/Vocabulary | Anterior Left | F(1, 180.91) = 1.82, p = 1.00 | F(2.66, 180.91) = 11.56, p < .001 *** |
| | Anterior Midline | F(1.01, 180.91) = 9.32, p = 0.04 * | F(2.73, 180.91) = 19.14, p < .001 *** |
| | Anterior Right | F(1, 180.91) = 6.61, p = 0.182 | F(2.74, 180.91) = 6.45, p = 0.003 ** |
| | Central Midline | F(1, 180.91) = 4.18, p = 0.734 | F(2.46, 180.91) = 11.13, p < .001 *** |
| Condition × ROI × OSpan | Central Left | F(1, 180.91) = 0.04, p = 1.00 | F(2.59, 180.91) = 6.38, p = 0.005 ** |
| Condition × ROI × Speaking/Grammar | Anterior Left | F(2.92, 180.91) = 25.14, p < .001 *** | F(1, 180.91) = 0.99, p = 1.00 |
| | Anterior Midline | F(2.89, 180.91) = 9.5, p < .001 *** | F(1, 180.91) = 4.64, p = 0.561 |
| | Anterior Right | F(2.76, 180.91) = 8.61, p < .001 *** | F(1, 180.91) = 1.76, p = 1.00 |

### 5.3.6. Phrase Structure Violations, 600-800 ms

The model describing responses to phrase structure violations at 600-800 ms, which was arrived at through our AIC-driven term selection, was:

*Voltage ~ Violation:ROI:**OSpan** + Violation:ROI:**Speaking/Grammar** + Violation:ROI:**Listening/Grammar** + Violation:ROI:**Listening/Vocabulary** + Violation:ROI + Violation + ROI + (1 + ROI | Participant)*

The model included a significant effect of condition ($F(1, 173 = 16.82$, $p < .001$), ROI ($F(8, 173=2.33$, $p = .017$), and their interaction ($F(8, 173) = 64.99$, $p < .001$). Post-hoc comparisons of the two-way condition by ROI interaction revealed that, across participants, the condition contrast was significant at all ROIs, where significance was corrected for 9 comparisons using a Bonferroni adjustment. This violation effect showed significant negativity at all ROIs except for the anterior left ROI, which showed significant positivity. The distribution of this negative response was consistent with that of the P600, which phrase structure violations were expected to elicit.

The three-way interactions between condition, ROI and each of the four included ID measures revealed that each influenced P600 amplitude in the present data. P600 amplitude was strongest in participants with lower Listening/Grammar scores for all ROIs, with the majority of variability in response amplitude being seen in lower-scoring individuals. Similarly, P600 amplitude was highest for participants with lower Listening/Vocabulary scores at central midline and posterior midline ROIs, with no overall trend seen in other ROIs. These effects can be seen in **Figure** 5.8 and **Table** 5.5.

Participants with higher Speaking/Grammar scores showed a higher-amplitude P600 at the central midline ROI. While this interaction depicted a degree of nonlinearity in posterior ROIs

and a strong sinusoidal fluctuation in anterior ROIs, no overall trend was evident, as the 95% CI of the contrast at these ROIs contained zero µV at various points on the ID measure spectrum.

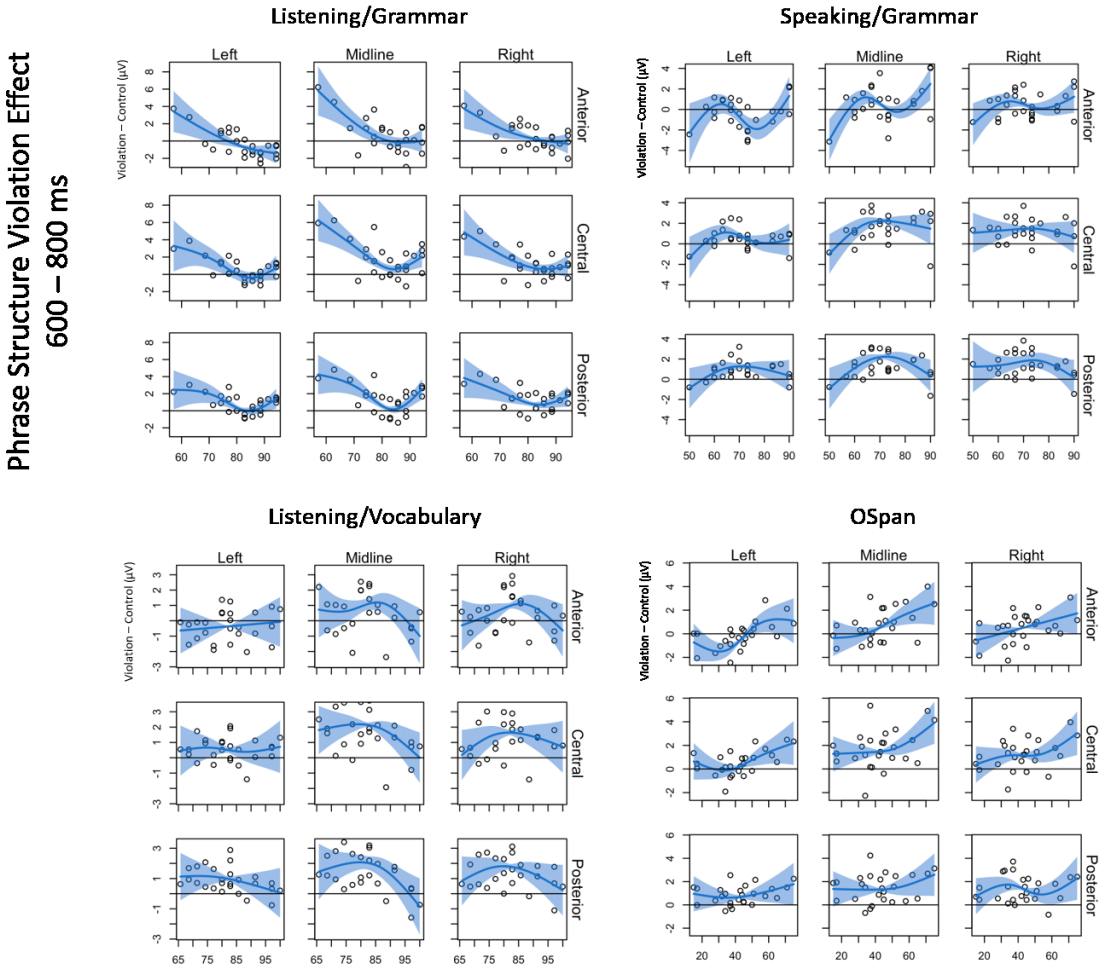


***Figure 5.8*** *Phrase structure violation effects in the 600-800 ms time window. The combined 95% confidence intervals of the sentence type estimates are shown, indicating where the two significantly diverge.*

***Table*** *5.5 List of condition effects (violation vs. control) at each ROI across participants, and significant three-way interactions between condition, ROI, and ID measures. Effects are limited to phrase structure violations in the 600-800 ms time window.*

| Interaction | ROI | Violation – Control Contrast | |
|---|---|---|---|
| Condition × ROI | Anterior Left | t(205110) = -4.1, p < .001, *** | |
| | Anterior Midline | t(205110) = 8.6, p < .001, *** | |
| | Anterior Right | t(205110) = 5.55, p < .001, *** | |
| | Central Left | t(205110) = 6.79, p < .001, *** | |
| | Central Midline | t(205110) = 26.31, p < .001, *** | |
| | Central Right | t(205110) = 15.72, p < .001, *** | |
| | Posterior Left | t(205110) = 10.75, p < .001, *** | |
| | Posterior Midline | t(205110) = 18.48, p < .001, *** | |
| | Posterior Right | t(205110) = 17.3, p < .001, *** | |
| **Interaction** | **ROI** | **Control Sentences** | **Violation Sentences** |
| Condition × ROI × Listening/Grammar | Anterior Left | F(2.27, 173.46) = 6.77, p = 0.02 * | F(1, 173.46) = 9.24, p = 0.042 * |
| | Anterior Midline | F(2.38, 173.46) = 4.07, p = 0.143 | F(1.97, 173.46) = 9.69, p = 0.001 *** |
| | Anterior Right | F(2.28, 173.46) = 5.23, p = 0.18 | F(1, 173.46) = 9.62, p = 0.035 * |
| | Central Midline | F(1.23, 173.46) = 2.5, p = 1.00 | F(2.89, 173.46) = 26.28, p < .001 *** |
| | Central Right | F(1.63, 173.46) = 1.53, p = 1.00 | F(2.41, 173.46) = 5.29, p = 0.029 * |
| | Posterior Left | F(1, 173.46) = 1.15, p = 1.00 | F(2.89, 173.46) = 11.67, p < .001 *** |
| | Posterior Midline | F(1, 173.46) = 3.53, p = 1.00 | F(2.94, 173.46) = 30.84, p < .001 *** |
| | Posterior Right | F(1, 173.46) = 0.69, p = 1.00 | F(2.76, 173.46) = 10.57, p < .001 *** |
| Condition × ROI × Listening/Vocabulary | Anterior Midline | F(1.24, 173.46) = 4.77, p = 0.298 | F(2.79, 173.46) = 10.4, p < .001 *** |
| | Anterior Right | F(1, 173.46) = 3.51, p = 1.00 | F(2.7, 173.46) = 10.93, p < .001 *** |
| | Central Midline | F(1, 173.46) = 3.25, p = 1.00 | F(2.63, 173.46) = 16.56, p < .001 *** |
| | Central Right | F(1, 173.46) = 2.28, p = 1.00 | F(2.58, 173.46) = 9.2, p < .001 *** |
| | Posterior Midline | F(1.42, 173.46) = 1.31, p = 1.00 | F(2.58, 173.46) = 6.93, p = 0.003 ** |
| | Posterior Right | F(1, 173.46) = 0.09, p = 1.00 | F(2.55, 173.46) = 8.91, p < .001 *** |
| Condition × ROI × OSpan | Anterior Left | F(2.9, 173.46) = 8.52, p < .001 *** | F(1, 173.46) = 4.97, p = 0.462 |
| | Central Left | F(1, 173.46) = 2.1, p = 1.00 | F(2.72, 173.46) = 8.12, p = 0.001 ** |
| | Central Midline | F(1, 173.46) = 4.12, p = 0.76 | F(2.65, 173.46) = 10.24, p = 0.003 ** |
| Condition × ROI × Speaking/Grammar | Anterior Left | F(1.99, 173.46) = 1.21, p = 1.00 | F(2.94, 173.46) = 15.41, p < .001 *** |
| | Anterior Midline | F(1, 173.46) = 0.33, p = 1.00 | F(2.97, 173.46) = 32.15, p < .001 *** |
| | Central Left | F(1, 173.46) = 0.02, p = 1.00 | F(2.88, 173.46) = 8.91, p = 0.001 *** |
| | Central Midline | F(1.01, 173.46) = 0.27, p = 1.00 | F(2.81, 173.46) = 16.16, p < .001 *** |
| | Posterior Left | F(1, 173.46) = 2.3, p = 1.00 | F(2.63, 173.46) = 8.05, p = 0.004 ** |
| | Posterior Midline | F(1, 173.46) = 2.48, p = 1.00 | F(2.78, 173.46) = 16.95, p < .001 *** |

### 5.3.7. Linear and Nonlinear Model Fit

The ability of the two modeling techniques (LME and GAMM) to describe observed variance was measured using marginal $R^2$, which indicates the proportion of variance accounted for by the fixed effects of a model and that is not explained by a trivial model applied to the same dataset (i.e., a model with zero predicted response amplitude for each predictor) (Nakagawa & Schielzeth, 2013). This measure was calculated for each sentence type and time window. In addition to $R^2$, which provided an indication of model fit to the observed data, we also used the AIC to evaluate the trade-off between model parsimony and log-likelihood. Given that LME and GAMM models were built using the same data set, the AIC associated with each model was considered comparable between the two.

All $R^2$ estimates are depicted in **Table 5.6**. On average, LME models had an $R^2$ of 2.81% (SD = 0.86%). GAMM models described considerably more variance, with a mean $R^2$ of 7.01% (SD = 0.19%). In this instance, GAMM explains significantly more variance than LME in comparison with their trivial models (t(4.01) = 3.89, *p* = .017).

**Table 5.6** Marginal $R^2$ for each model, indicating accounted variance that is not explained by a trivial model. Fit is shown for each of the four models for each modeling technique (LME and GAMM).

| | | 300-500 MS | | 600-800 MS | |
|---|---|---|---|---|---|
| | | **LME** | **GAMM** | **LME** | **GAMM** |
| **$R^2$** | Semantic | 3.13% | 8.94% | 3.83% | 8.40% |
| | Phrase structure | 1.84% | 5.91% | 2.45% | 4.80% |

The difference in model fit provided by the two techniques was considerable. Overall (i.e., across techniques), model fit did not differ significantly between semantic ($R^2$ = 6.07%, SD =

3.01%) or phrase structure ($R^2$ = 3.74%, SD = 1.92%) violation effects (t(5.09) = 1.3, $p$ = 0.25). Similarly, no significant difference was found between fit of models in the 300-500 ms ($R^2$ = 4.95%, SD = 3.15%) and 600-800 ms ($R^2$ = 4.86%, SD = 2.54%) time windows (t(5.74) = 0.04, $p$ = 0.96). There was no significant difference in model fit between time windows, with an overall average $R^2$ of 4.91% (SD = 2.65%). These measures are averaged across modeling techniques. These differences are noted in **Table 5.6**.

A notable difference between the two techniques is that, due to the ability of GAMM to account for more variance with each added term than LME in general resulting from the flexibility of nonlinear terms in reducing residual variance, the addition of these terms was more often considered an improvement for the model despite penalties for their addition (i.e., improved AIC). The result was that, while predictors were frequently disregarded when using LME due to their improvement not being strong enough to outweigh this penalty, this was rarely the case when using GAMM. Each of the GAMM models built included the maximum number of four ID measures. Conversely, LME models included an average of 2.89 predictors per model. Note that of the seven predictors investigated, only a maximum of four could be included in any one model due to their collinearity between three pairs, described in the collinearity reduction process of Chapter 2.

Importantly, predictors were included on the basis of improved log-likelihood. However, it is possible that a predictor can improve model likelihood while not contributing significant effects or interactions. It is important to note that due to differences in the mechanics of the two modeling techniques, a direct comparison of sensitivity to effects at specific regions is not possible. As discussed above, LME natively estimates the significance of contrasts, including the violation effect of each sentence type. This is true not only in the contrast of condition as a main effect of the predictor, but also in interactions with an ID measure at an ROI. Conversely, only

the significance of individual smooth terms can be estimated using GAMM. For this reason, the closest comparison of significance in violation effects between the two techniques was through comparing their 95% CI. Again, while LME represents the 95% CI of the contrast term, this must be estimated through GAMM as described above. Therefore, while the frequency of effect significance at specific ROIs would not be a meaningful comparison, fewer ROIs showed significant differences between control and violation conditions when investigated using GAMM. Even so, without knowledge of the true violation effect at each ROI, it is not possible to determine whether finding the violation effect to be significant at more ROIs using LME reflects improved sensitivity to underlying effects, or false positive findings.

### 5.3.8. Smooth Fits to Random Variance

While nonlinearity in the interactions between ID measures and violation effect size was predicted, the strong and repetitive fluctuations depicted by GAMM in some cases were beyond what could reasonably be expected, or what have been suggested in our previous findings in Chapter 3 and Chapter 4. For example, in the case of Speaking/Grammar, increasing scores were associated with an alternating significant negative and positive violation effect at the front midline ROI as scores increased (**Figure** 5.8). The strong adherence of the smooth interaction term fit to the observed data suggested that this pattern may have resulted from overfitting, depicting significant effects where a condition contrast was only shown in a few participants. In cases such as these, this may have been caused by small sample size relative to the degree of variability in violation effect size. This prompted an investigation into the effect of sample size on model fit under the present distribution of responses. This was of particular concern in the three-way interaction between condition, ROI, and Speaking/Grammar 600-800 ms following the presentation of phrase structure violations, where over-fitting was suspected.

A random normal distribution of scalp voltages were generated for each condition (violation and control sentences) at each ROI, adhering to the mean and standard deviation of responses observed for each conditions and ROI across participants. This ensured that the violation – control conditional contrast at each ROI reflected that seen in the real data, while randomly assigning response amplitudes to observed ID measure scores ensured that there should be no systematic influence of any ID measure on the violation effect in these simulated data. Rather, any apparent influence of ID measures should have resulted from chance alone. The distribution of responses to phrase structure violations during the 600-800 ms time window were used as our basis for producing simulated data, and accordingly a model structure identical to that which described this condition and time window above was used.

As the sinusoidal fit seen in several ROIs for the influence of Speaking/Grammar was our primary concern, one representative smooth fit of the influence of Speaking/Grammar on violation effect amplitude was produced for each of twenty sets of randomly generated data. Ten of these sets included 33 participants, mirroring the present data, and ten included 66 participants to investigate what impact doubling our sample size might have. While results are shown for Speaking/Grammar scores alone, as this is the ID measure which was associated with sinusoidal fluctuations in violation effect amplitude, as described the model structure was identical to that used to depict responses to phrase structure violations in the 600-800 ms time window. One representative fit to a simulated data set of each sample size are shown in **Figure 5.9**, though ten identical models were fit to randomized data sets of each size.

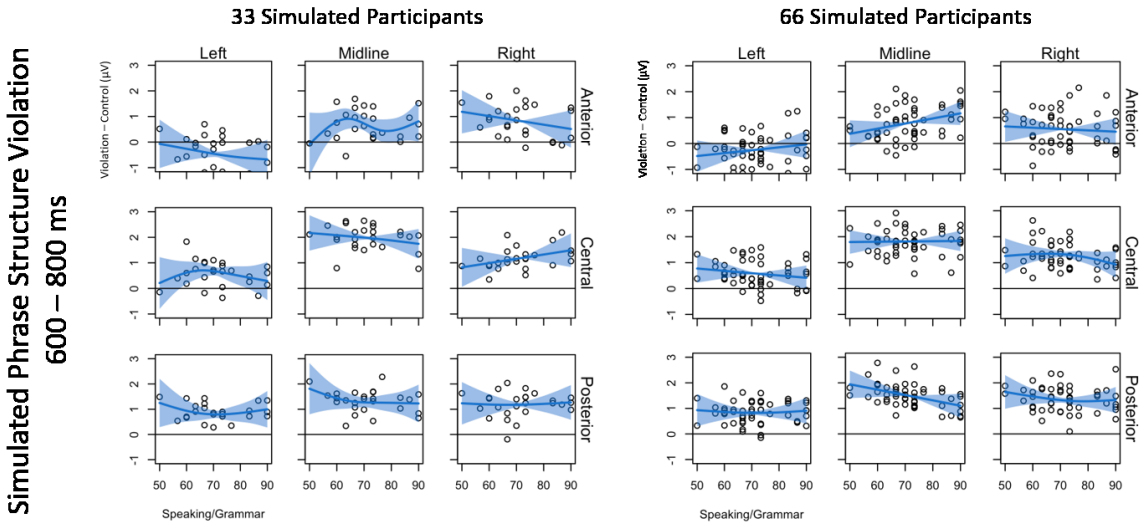


**Figure 5.9** Simulated data following the mean and standard deviation response amplitude for each condition in each ROI, averaged across participants to ensure that any influence of Speaking/Grammar scores is due to chance. Data are shown for 33 simulated participants (left), equivalent to the present data set, and for 66 simulated participants (right), to represent a doubling of our present sample size.

Models fit to simulated data sets of each sample size were largely consistent in their characteristics across randomized iterations. Overall violation effect amplitude at each ROI was typically similar regardless of sample size, while 95% confidence intervals narrowed with additional data. However, we were primarily interested in the propensity for producing s-shaped curves or sinusoidal fits in each of the two sample sizes, as this type of relationship should reflect fitting to random variance in which no systematic effect should exist. To investigate this we investigated two measures in each Speaking/Grammar smooth fit. First, we investigated the number of ROIs in each of the ten simulations per sample size which showed evidence for a false positive influence of Speaking/Grammar. That is, data were designed to contain no systematic influence of Speaking/Grammar scores (or any other ID measure) on violation effect amplitude, as amplitude was randomized across participants while maintaining the distribution of Speaking/Grammar scores seen in the original data. Therefore any influence of

Speaking/Grammar could be concluded to be a false positive. To this end, we operationalized a false positive as being any single ROI at which a line with zero slope cannot be drawn from the lowest to the highest Speaking/Grammar scores while being contained entirely within the 95% confidence interval of the smooth fit.

Second, as any relationship between Speaking/Grammar scores and violation effect amplitude resulted from chance alone, nonlinear smooth fits of Speaking/Grammar scores to responses were concluded to reflect over-fitting. We therefore investigated the number of individual ROIs at which a parabolic smooth fit had been depicted. The number of ROIs showing either of these two characteristics across the ten simulations are shown in **Table** 5.7 for models fit to data sets of each size (33 and 66 participants).

**Table** 5.7 The number of ROIs (out of a possible 9) showing either a false positive influence of Speaking/Grammar score on violation effect amplitude (linear or nonlinear), or a parabolic smooth fit of Speaking/Grammar score influence on violation effect amplitude, where simulated data were designed to have no systematic influence of Speaking/Grammar. Results are shown for simulated data sets that include either 33 participants (reflecting the present sample size), or 66 participants (a doubling of the present sample size).

| | **False Positives** | | **Parabolic Functions** | |
|---|---|---|---|---|
| | n=33 | n=66 | n=33 | n=66 |
| Simulation 1 | 0 | 1 | 2 | 1 |
| Simulation 2 | 0 | 1 | 2 | 1 |
| Simulation 3 | 0 | 0 | 2 | 1 |
| Simulation 4 | 0 | 0 | 3 | 2 |
| Simulation 5 | 1 | 1 | 1 | 2 |
| Simulation 6 | 0 | 0 | 3 | 0 |
| Simulation 7 | 0 | 0 | 3 | 1 |
| Simulation 8 | 0 | 1 | 3 | 1 |
| Simulation 9 | 0 | 0 | 4 | 2 |
| Simulation 10 | 2 | 1 | 5 | 1 |
| **Mean** | 0.30 | 0.50 | 2.80 | 1.20 |
| **95% CI** | 0.42 | 0.33 | 0.70 | 0.39 |

Overall, the number of false positives was similar regardless of sample size, though a slight increase in this number was evident for fits to 66 participants (0.50 ROIs per simulation on average showing false positive influence of Speaking/Grammar) than 33 participants (0.30 ROIs per simulation). This was likely related to the characteristic narrowing of 95% confidence intervals that was associated with increasing sample size, which would result in a zero-slope line being less likely to cross the Speaking/Grammar score spectrum entirely within the 95% confidence intervals of the fit. Conversely, increasing the sample size (provided data contain no effect) might also be expected to reduce spurious Speaking/Grammar influence overall, offsetting this increase in error. Counterbalancing of these two effects may have resulted in the overlap in the 95% confidence intervals surrounding the mean number of ROIs showing false positives for each sample size, as noted in **Table** 5.7, and diminishing any difference overall between the two sample sizes.

Interestingly, there was a considerable decrease in the number of ROIs at which a parabolic smooth fit was depicted as the sample size increased, from a mean of 2.80 ROIs per simulation when modeling data for 33 participants to 1.20 ROIs for 66 participants. Moreover, as noted in **Table** 5.7, the 95% confidence intervals surrounding these error counts are non-overlapping, suggesting that doubling of our present sample size might improve over-fitting to random individual variance in response amplitudes that may not be related to ID measures. Importantly, the assumption that a nonlinear fit directly reflects over-fitting is only valid in the context of these simulations, where it is known that nonlinear dependencies should not exist. Therefore, these findings can only speak to the propensity for GAMM to over-fit in the presence of a truly linear effect (or no effect at all), but cannot validate – or invalidate – fits to relationships which may truly be nonlinear in nature. Nonetheless, these results suggest that

over-fitting may be taking place at our present sample size of 33 participants, and that these concerns may be alleviated somewhat at larger sample sizes.

## 5.4. Discussion

### 5.4.1. Overview of Objectives

The present study aimed to optimize the current standard practices in statistical modeling for individual differences in language processing through relaxing the assumption of linearity in the relationship between ID measures and cortical responses to language violations. Expanding on our previous investigations of these relationships, which evaluated the dependence of responses to semantic or phrase structure violations on various ID measures using linear mixed models, the present analyses used generalized additive mixed models in an attempt to improve model fit, likelihood, and sensitivity to effects. ID measures of interest included aspects of grammatical ability, vocabulary, reading efficiency, speech comprehension and working memory capacity. These analyses were performed on the same data set as was used for our linear mixed modeling investigations, in order to compare metrics of model competence as directly as possible.

### 5.4.2. Choice of Regression Spline Type

Despite the considerable mechanical differences in the model-building process whether using restricted cubic or thin plate regression splines, the impact on model fit was negligible. When using the same formula and set of observations, fits achieved using the two methods were visually indistinguishable, and resulted in no differentiation between models in terms of the amplitude of predicted responses, confidence interval width, residual variance or significance of terms. Moreover, AIC differences between the two barely met the criteria of a

worthwhile improvement, with a difference of 14 between models using the two spline types (K P Burnham & Anderson, 2004). Nonetheless, as a choice was required, all final models used cubic regression splines due to their marginal improvement (in terms of AIC) over thin plate regression splines. Given these findings, the choice of spline type in the present data was not deemed to have an impact on model fit overall. Notably, the two splines types were only evaluated using 4 knots (cubic regression splines) and 4 basis functions (thin-plate regression splines). The possibility remains that one spline type might emerge as preferable in more complex models (i.e., more knots or basis functions), but our preference toward using a simplistic model to mirror the possible shapes of functions suggested in Chapter 3 precluded investigation of their efficacy in such circumstances, as this would not be applicable to the present research question.

### 5.4.3. Fit Complexity

The more likely parameter to be impacted by data set size was the specification of maximum model complexity (i.e., the number of knots in a smooth term). Keep in mind that our discussions of complexity refer to the potential for intermittent variability in response amplitude across an ID measure's spectrum — that is, the 'curviness' of each spline rather than the overall number of smooth terms. While specifying a maximum of ten knots may still result in a linear fit, doing so also allows for the possibility to produce an erratically curvy fit for a data set of insufficient size to support such complexity. Conversely, this degree of curviness cannot be the case when using only two or three knots. As the maximum possible complexity of that a model can support is a function of the data set size, the ideal complexity specification for a study, which is in part determined by the experimenter (maximum complexity) and in part by *gam* using generalized cross validation, is unique to the sample being investigated and the expected effect

size. Therefore, general guidelines to this specification cannot be established. Instead, expectations based on prior research and preliminary results might provide the best guidance when selecting the maximum complexity.

Evaluating the appropriateness of varying degrees of complexity presented a number of challenges. GAMM uses generalized cross-validation to apply unique quadratic penalties to individual splines, attempting to trade maximum descriptive ability for model parsimony, similar to an optimization of model terms using AIC (Wood, 2006; Wood, 2011). While most of the time individual observations were not found to unduly influence model fit, this was clearly not the case in all interactions. Moreover, in the absence of a functioning limitation on model complexity, the AIC was not found to be appropriate for evaluating the number of knots. While the AIC has proven effective in limiting the number of terms in a model, a single smooth term can be composed of any number of splines, and additional splines are not penalized. The result was that the model with more splines invariably achieved a better fit to the observed data and was assigned 100% conditional probability when using AIC weights (i.e., AIC weight of 1.0).

Given these challenges, empirically determining the ideal degree of complexity for a model using the present data set was not possible. Qualitatively, allowing for four knots (three splines) resulted in intermittent fluctuations in violation effect size that coincided with only slight changes in ID measures, and this problem was exacerbated at five or six knots. Again, these fluctuations were not supported by the motivating research. Even while some interactions demonstrated these fluctuations with only four knots (allowing for three cubic regression splines), the results overall did not appear unduly influenced by individual responses. Moreover, we felt that allowing for up to four knots could provide a means to model the maximum expected degree of complexity based on our previous observations (i.e., parabolic effect, with potential for a linear segment, as suggested in Chapter 3), However, this type of qualitative

determinant of an appropriate degree of model complexity was not ideal as fits appeared to conform strongly to the responses of a small number of individuals in several interactions, and our concern was that findings of this nature may not generalize.

### 5.4.4. Reliability of Nonlinear Interactions

It was hypothesized that GAMM could be used to describe nonlinearities in the dependency of the violation effects on ID measures which linear modeling solutions would be incapable of describing. While subtle changes in model fit that improve accuracy and sensitivity are an obvious advantage, understanding nonlinear dependencies may be an important part of elucidating the relationship between language proficiency and processing. It is important to note that, while not always the case, most of the interactions depicted did not suggest that the observed nonlinearities were meaningful or replicable. That is, while many interactions between violation effects and ID measures were nonlinear, the confidence intervals for these effects frequently did not diverge from zero, or could contain a zero-slope line. In addition, violation effects that showed a sinusoidal dependency on an ID measure, but no overall trend, may similarly reflect that a violation effect is not dependent on an ID measure alongside over-fitting. In short, neither type of interaction may be replicable in other samples.

### 5.4.5. Comparison of Findings with LME and GAMM

When evaluating the overall findings represented through LME and GAMM, in terms of the presence and directionality of ID measures' influence on violation effect size, similarities were noted. For example, the time period of 600-800 ms following the onset of phrase structure violations was revealed an influence of Listening/Grammar scores on response amplitude that was similarly modeled by both LME and GAMM. Notably, both techniques suggested that lower

Listening/Grammar scores were associated with a stronger P600 response, as outlined in **Figure 5.10**.

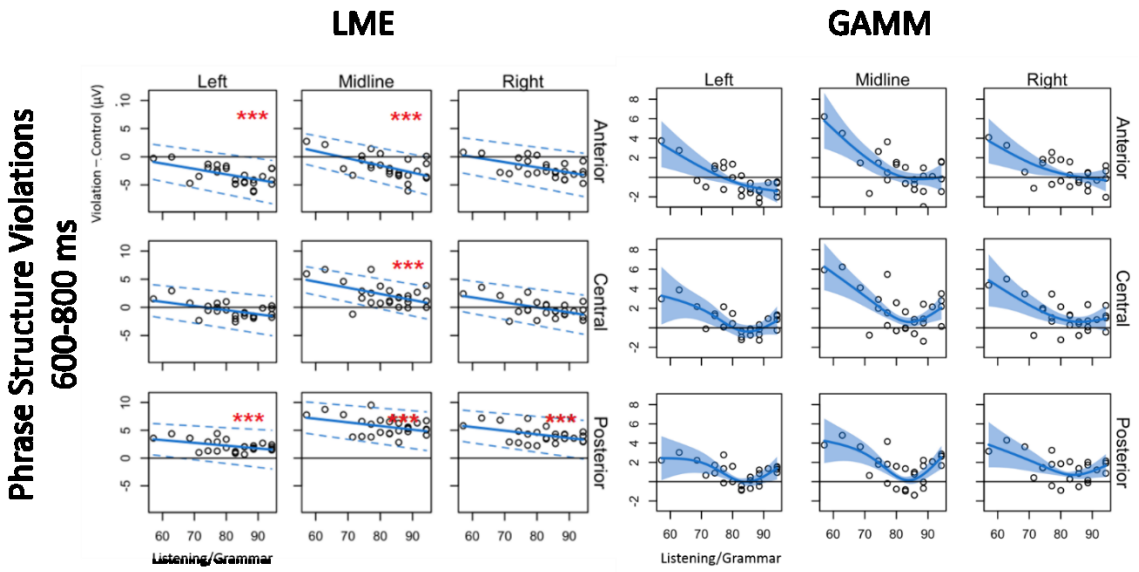


**Figure 5.10** Violation – control response contrast to phrase structure violations in the 600-800 ms time window, modeled using LME (left) and GAMM (right).

While this relationship was identified as significant at the six of nine ROIs using LME, it was significant at all ROIs using GAMM. This is noteworthy as, not only did GAMM depict a nonlinear interaction between violation effect size and Listening/Grammar score, this relationship suggested that the strongest (and potentially least linear) effects were demonstrated at frontal ROIs. This may be a case where the presence of an effect only in those participants with the lowest Listening/Grammar scores resulted in LME missing the effect altogether at several ROIs. However, due to our concerns regarding potential over-fitting, it cannot be known whether this fit resulted from the influence of the small number of lower-scoring individuals.

During this same time period, a similar overall trend was seen across other ID measures of interest when comparing the two techniques. However, GAMM often estimated the

amplitude of responses within three-way interactions (condition, ROI, and ID measures) to be of much lower amplitude than did LME. The same could be said for the time window of 300-500 ms following the onset of semantic violations, as the same ID measures were seen to affect N400 amplitude, and a similar overall trend was seen when comparing the two techniques. However, in this time window, GAMM did not identify any reliable nonlinear interactions. Interactions identified using LME and GAMM therefore appeared to reflect the same overall trends across sentence types and time windows, while nonlinearities depicted using GAMM within those trends appeared to be strongly influenced by a small number of participants in most cases.

### 5.4.6. Evaluating Model Fit

It was predicted that the use of GAMM would result in a model fit that more accurately depicts the relationship between ID measures and cortical responses than was achieved using LME. This was expected due to the ability of GAMM to produce nonlinear fits of dependent to independent variables. First, this was evaluated using $R^2$, which can be considered a comparison of the model-predicted values with those observed. This is an important metric in the ability of the model to depict dependencies between the dependent variable and predictors, but cannot be considered an overall measure of 'accuracy' for several reasons. Ideally, a model strives to perfectly describe the observed variance (i.e., scalp voltage in response to language violations or well-formed sentences) using a set of predictors, while also being generalizable to a population. However, the two motives are often in opposition. Building a model with enough complexity will inevitably account for all participant- or group-specific deviations from the norm in a response, but such a model can be overly-specific to the described group, to the point that the added complexity may not be reliable in other samples.

This pattern of over-fitting can be lessened (but not mitigated entirely) by limiting the model to only include variables that were hypothesized *a priori* to have some influence on a dependent variable, as well as limiting the degree of complexity and nuance that is allowed in those relationships. However, over-fitting is an ever-present concern. Particularly in an exploratory investigation into methods of improving model fit, the limited usefulness of $R^2$ alone must be considered, and metrics accounting for model complexity should be used in conjunction.

Assessments of model fit can therefore be improved through consideration of the AIC, which optimizes model parsimony – a trade-off between model likelihood and complexity. Indeed, hypotheses were confirmed as $R^2$ showed a considerable improvement in models generated using GAMM over those built using LME, doubling or even nearly tripling in value. This was reflected in a strong improvement in raw AIC value, and in all cases 100% conditional probability assigned to models generated using GAMM by AIC weights (i.e., AIC weights of 1.0).

While the improvement in model fit is unsurprising, given that a transition to nonlinear terms should be expected to result in reduced residuals and therefore increased $R^2$ at no extra cost to model complexity when compared with LME (Sánchez, 1982), the magnitude of AIC improvement is noteworthy as it suggests a considerable and worthwhile improvement over LME while using the same guidelines established in our previous chapter. Moreover, recall that residual variance is akin to a model's error term, and so reduced residuals necessarily improve the *F* statistic of any term's significance test or related *post-hoc* testing. Therefore, the demonstrated improvement in model fit also translates into improved ability to determine significance of effects (i.e., improved statistical power). This can be advantageous in investigations into small, elusive or inconsistent effects.

It should be noted that while using the AIC solves the problem of model complexity that $R^2$ alone does not address, it remains unable to predict reliability of the model. That is, whether these effects generalize to a population. In this sense replication is the only true measure of reliability, through application of the model to novel data, making continued testing of the effects presently described imperative. Despite this, the present results support the hypothesis that when provided identical input, GAMM produced considerably more descriptive models that demonstrated a notable improvement in model fit and likelihood when compared with LME.

### 5.4.7. Sample Size Simulations

The tendency for nonlinear deviations in an interaction to be guided by small numbers (often two or three) of participants suggested that the improvements in model fit afforded using GAMM also came at the detriment of over-fitting. For example, increases in Speaking/Grammar scores (as well as other ID measures) were associated with sinusoidal fluctuations in phrase structure violation amplitude during the 600-800 ms time window. Given that our findings in Chapter 3 and Chapter 4 suggested that the influence of these ID measures might be best described using linear, asymptotic, or at the most complex parabolic functions, this type of sinusoidal fluctuation may have indicated that these smooth terms were conforming too closely to individual variance which was not reflected in an overall trend across participants.

The fit of smooth terms which appeared more complex than our raw data suggested should be necessary prompted an investigation into the propensity for the *gam* function to conform to random variability between participants despite no overall influence of an ID measure. To establish this, we designed a series of simulated data sets which mirrored our own data in all attributes, with the exception that random generation of responses was performed with no knowledge of specific ID measure scores, and so no systematic influence of any ID

measure should exist. Even under these conditions, fits to every simulated data set (using a model structure identical to that which was used in characterizing ID measure influence) consistently produced confidence intervals which could not contain zero-slope lines (suggesting a false positive ID measure influence) as well as parabolic functions (nonlinearities where no association should exist).

These findings suggested that, given the sample size and degree of variance in our data, similar patterns of conforming to random individual-level variance to produce nonlinear fits and significant estimates may have occurred. Interestingly, however, the number of ROIs showing nonlinear fits was significantly reduced as we doubled the sample size of our simulated data sets, suggesting that these problems stemmed in part from the number of participants in our data set. It is important to note that the number of participants required to mitigate over-fitting in our simulations was specific to the variance and ID measure distribution of the present data set. However, results suggest that the present data set was not well-suited to modeling nonlinear interactions of this nature using GAMM.

#### 5.4.8. Conclusions

GAMM demonstrated consistently superior model fit when compared with LME, both in terms of reduced residuals (higher $R^2$) and producing more parsimonious models (evaluated using AIC). Moreover, the general trend of the influence of ID measures mirrored that which was described using LME. However, within these trends, nonlinear fits appeared to conform strongly to the responses of small numbers of individuals, suggesting over-fitting. While the choice of spline type appeared to have negligible impact on model fit, the problem was exacerbated as models were permitted to compute additional splines, increasing the complexity of smooth fits. Moreover, nonlinear fits to random variability were identified in simulated data which contained

no effect of ID measures on response amplitude. These findings raised suspicion that improved model fit was achieved at the expense of over-fitting to random individual level variance, which may be mitigated at larger sample sizes.

# Chapter 6: Data-driven characterization of individual differences in language violation processing using CForest

## 6.1. Introduction

In previous chapters, we have demonstrated evidence to support that individual differences in language proficiency and other cognitive factors are related to violation effect amplitude and distribution, and potentially the strategies that are used for language processing (Liang and Chen, 2014; Moreno and Kutas, 2005; Pakulak and Neville, 2010; Tanner, 2013; Tanner and Van Hell, 2014; Weber-Fox et al., 2003). Using electroencephalography (EEG), we have demonstrated that differences in the ability to discern a sentence's grammatical structure as well as in vocabulary size may impact two important components of event-related potentials (ERPs) that are frequently used to index the learning and processing of various aspects of language. Not only do these IDs affect ERPs associated with language processing, but working memory capacity has shown a similar dependence, replicating what has been suggested regarding its effect on the P600 (Nakano et al., 2010). These findings highlight the need to consider these types of ID measures in investigations of language processing, even if only as a covariate where the effects are not pertinent to the research question.

We have shown that allowing for nonlinearities in the relationships between ID measure scores and ERP component amplitude may result in improved model fit, but may also require a larger sample size than linear modeling solutions. One result was considerable improvement in model fit ($R^2$), and therefore reduced residuals. Note that improved fit alone does not speak to generalizability of a model, as addition of terms until a term exists for each observation will necessarily drive $R^2$ toward 1.00 (i.e., over-fitting). However, terms were added in consideration

of the AIC, which applies penalties for the number of terms against the likelihood of the model (Akaike, 1974). Penalties are linear in relation to the number of terms in a model. In addition to fit, however, allowing for nonlinearities described important nuances in the relationships between certain ID measures and violation effects, such as identifying ranges of ID measure scores across which changes in the violation effect (violation – control sentence response amplitude) were most evident. In many cases, the range of variability in violation effect amplitude was restricted to a portion of the ID measure spectrum. It was not surprising that these nonlinearities were present, given that it is unlikely for cortical response amplitude to change in one-to-one correspondence with testing scores.

The procedures outlined above have added to the mounting evidence that proficiency-related effects are likely ubiquitous in sentence-processing tasks, and have provided some means of improving sensitivity to those relationships. However, several deficiencies must still be considered. Most notably, regression models commonly require adjustment for multiple comparisons, and to minimize the number of comparisons, EEG electrodes are frequently grouped into regions of interest (ROIs) based on spatial proximity. The work described above used a pre-determined 3x3 grid of ROIs, based on a roughly equal division of electrodes into scalp regions. This approach assumes that an effect will ideally be visible at all electrodes in the ROI, and if the assumption is not met, there is a corresponding loss of sensitivity as the effect is diminished by electrodes at which it is not exhibited. Moreover, evaluations of topography based on ROIs alone are coarse. While topography plots can be evaluated regarding the shape and extent of effects, and qualitative comparisons can be drawn, it can be difficult to know whether topographical differences between groups or conditions are reliable.

As an alternative to evaluating the significance of effect topographies using ROIs, clustering solutions can be used to group electrodes based on similarity of response amplitude

(Pernet et al., 2015a). One limitation to this approach, however, is that ID measure score bins (e.g., low- or high-proficiency) must be determined beforehand, and interactions between proficiency and response topography can be difficult to ascertain. The solution to this problem should be a form of significance testing that evaluates individual electrodes without requiring a correction for contrasts at every permutation of electrode groupings, while allowing for interactions between response topography other variables.

Also worth considering is that heteroscedasticity proved problematic in some interactions identified using GAMM, where strong influence by small numbers of individuals resulted in effects that were not supported by the motivating literature and were likely not replicable (Liang and Chen, 2014; Moreno and Kutas, 2005; Pakulak and Neville, 2010; Tanner, 2013; Tanner and Van Hell, 2014; Weber-Fox et al., 2003). While potentially important nonlinearities were identified (alongside considerable improvements in model fit and parsimony), these findings suggested that a nonparametric approach may be required to appropriately characterize proficiency effects given the sample size.

Taking advantage of developments in machine learning may provide steps toward solutions for each of the above concerns, resulting in an overall improvement of sensitivity, and a more data-driven characterization of effects. Recently, a combination of machine learning procedures known as conditional inference random forest analysis (CForest) has been applied for nonlinear characterization of white matter integrity using diffusion tensor imaging (DTI) data, providing a robust detection of subtle and often-elusive effects related to sex and age (McWhinney et al., 2016). CForest belongs to a group of recursive binary partitioning techniques, detailed by Hothorn et al. (2006), and refers to a specific implementation of random forest analysis, which will be described in greater detail below.

Using this technique, a response variable (e.g., cortical response amplitude for either violation or control sentences) is modeled as a function of some number of predictor variables. Conventionally, regression uses an additive approach, which linearly combines terms multiplied by coefficients to produce estimated responses, where each term represents the effect of one predictor or interactions between predictors on a response variable. CForest instead iteratively subdivides a data set into groupings of two opposing sets of values on a predictor, for example separating a data set into one containing only males and another containing only females, or similarly one containing younger and another containing older participants. Relationships between a predictor and the response variable are determined using a permutation testing framework, which maximizes the absolute value of the test statistic to determine the ideal 'split point' on a variable, which is then used to subdivide the partition into two smaller partitions (Strasser & Weber, 1999). This process is recursively performed on increasingly small subsets of data until either no further significant associations can be detected, some minimum number of observations exists in a partition, or a maximum number of divisions have been created. The result is a branching tree-like structure, where each split represents a significant difference between the response variables of the two predictor groupings. Predictors can include continuous variables (e.g., age, or ID measures), or categorical variables such as sex of scalp electrode. In the event that a categorical variable has three or more levels, subdivisions can be created using any two arbitrary groupings of levels.

This recursive partitioning process is used to produce a single 'tree', which can be used to determine the estimated response variable for a participant. For example, consider a simple model of height as a function of sex and age. If this model identifies a difference first and foremost by age, and further in gender – but only in the subdivision that includes males – then this tree has three terminal branches: Females, young males, and old males. Determining an

estimate is therefore as simple as knowing the age and sex of a participant. In this example, we consider only a single division on the age predictor, but this is not necessary. Numerous divisions are possible within increasingly small subsets of the initial data set. Moreover, using this approach, permutation testing in each node (partition) of this tree is controlled for multiple comparisons. However, this only represents a single tree. A forest can consist of hundreds, or thousands of trees, each of which can produce a unique estimate. The result is a distribution of estimates, allowing the user to derive confidence intervals and perform hypothesis testing.

Using the approach as detailed above, multiple trees would be identical. Thus in random forest analysis, there are two ways that a degree of randomness is introduced, which provide several advantages. First, a tree is only created using a randomized subset of the data, and so as the number of observations increases, the number of possible randomized combinations of data samples (i.e., unique trees) increases exponentially. Creating trees using a random sample of the full data set allows the model to produce estimates which occasionally exclude outliers. This process is therefore similar to bootstrapping. This random exclusion of observations across a large number of trees has a similar effect to accounting for random effects when using LME or GAMM. That is, through computation of a large number of trees (where each tree corresponds to one set of estimated responses in a randomized subset of the data, and the distribution of responses can be used to derive confidence intervals), a number of these estimates necessarily exclude the influence of specific electrodes, trials, or participants. The distribution of estimates is therefore reflective of random variability associated with any of the predictors in the model (Grandvalet, 2004). Second, not all predictors are evaluated for subdivision at every branch in a tree. By randomly excluding a portion of the predictors at each node, we ensure that subdivisions are not dominated by the predictor(s) most related to the response variable. This

allows for detection of smaller, but significant, effects. These processes will be explained in more detail below, in the Methods.

Using this approach seamlessly accounts for interactions through data-driven detection of optimal split points. As described in our simple example model above, if age is the predictor with the strongest association to response, and a significant age division exists, two new subdivisions will result. If one of those subdivisions is significantly associated with (and divided by) gender, while the other is not, then we have characterized an interaction between age and gender. Importantly, this is done so in a way that does not require post-hoc testing or correction for multiple comparisons. Rather than performing a number of contrasts, we have instead deductively ascertained the presence of significant differences. We can simply describe the difference in the magnitude of the predicted response that was shown between male and female adults, where that subdivision was not seen in the younger participants. This branching set of permutation tests allows for complex and nuanced characterization of high-level interactions without the need for stringent corrections or *a priori* predictions. While multiple comparison correction is intended to control for inflation of Type I error when performing numerous hypothesis tests in a single sample, delineating patterns in distinct branches instead performs hypothesis testing in distinct samples. Moreover, both effect size and significance can be described in any branch.

The CForest approach offers several advantages over previous characterizations of proficiency-related effects that have used regression models. First, response estimates can be predicted for individual electrodes wherever significant differences exist, rather than grouping electrodes into *a priori* regions of interest (ROIs). Effect topographies can therefore more accurately depict the true scalp distributions of effects, rather than being forced through a coarse and *a priori* defined set of ROIs. The result is that the scalp distribution can be described

with statistical probability. This circumvents the need to draw qualitative conclusions regarding topographical differences between groups or conditions from more heuristic approaches. Second, significance testing in each tree node determines stopping points, which prevents over-fitting (Hothorn et al., 2006). A division is not made unless it is both statistically significant and exceeds a user-defined threshold for subset size.

Third, subsampling in each tree accounts for outliers (bagging; Strobl et al., 2009). As described above, each tree randomly excludes some portion of the data set. If the data set contains outliers, this ensures that a proportion of a forest's trees are created which are less influenced, or potentially not influenced at all, by these outliers. Observations are not necessarily excluded at the participant level, and so a subsample may include some portion of each participants' data. With enough trees, by random chance entire participants are likely to be excluded from some trees. Therefore, participants with average responses or ID measure scores that are strongly divergent from the average will be probabilistically represented in the forest based on the subsampling proportion and the number of trees. As a distribution of estimates is drawn from the ensemble of trees, this results in a distribution with varied degrees of influence from outliers. In addition, where regression estimates are strongly leveraged by extreme predictor values, these values will only strongly influence branches if they coincide with extreme response variable values.

Lastly, the variability of response estimates produced across trees is improved by randomizing the predictor variables that are evaluated at each branch of a tree. As described above, only the predictor that is most-strongly associated with response values in a partition is considered for guiding further subdivision of that partition. It is highly likely therefore that a single strongly-associated predictor will guide many or all subdivisions until a partition is too small to support further identification of associations. This could preclude identification of

associations and significant subdivisions in more weakly-associated predictor variables, if the variables considered were not randomized. Moreover, a large number of trees in the ensemble will show divisions dominated by the most strongly-associated predictors, and will therefore produce estimates that are more similar to one another. This problem is considerably alleviated by randomly excluding a proportion of predictor variables from consideration when identifying associations in a partition (Variable pre-selection; Strobl et al., 2009). The result is more sensitive detection of subtle effects, and more varied trees.

Taken together, CForest demonstrates a number of advantages over conventional regression approaches. Considering the often-elusive nature of proficiency- and cognition-related effects, as well as variability between studies, this non-parametric, data-driven approach may be better suited to investigations of ID measures in language processing. While CForest has frequently been used for classification problems (e.g., Parvinnia et al., 2014; Rashid et al., 2011; Zainuddin et al., 2012) in which associations are learned to predict a response variable in a novel combination of predictors, they can also be used for characterization of responses in order to contrast groups, conditions, times or regions in neuroimaging data (McWhinney et al., 2016). This framework was in part designed to address the "small *n* large *p*" problem of many classification studies (Hothorn et al., 2006) — more predictor variables than observation units — which is of immense value in genetics & bioinformatics (for a review of the rapidly developing field, see Libbrecht and Noble, 2015). However, the ability to consider a large number of variables without penalty remains beneficial for descriptive models as well. This allows conceptualization of investigations with finer detail and larger scope, while avoiding inflation of Type I error.

The present analysis used CForest to delineate relationships between several ID measures and the spatial and temporal characteristics of ERPs elicited in response to sentence

violations. This included differences in scalp topography, response amplitude, and latency associated with the ID measures, as in the previous chapters. Specifically, as in the previous LMER and GAMM analyses, we investigated how these factors were modulated by language proficiency (vocabulary, grammatical ability), speech comprehension, word reading efficiency, and working memory.

N400 and P600 responses to semantic and phrase structure violations, respectively, were first characterized using CForest to align subsequent analyses both with previous chapters and the motivating research (Nakano et al., 2010; Pakulak & Neville, 2010; Tanner, 2013; Tanner et al., 2014; Tanner & Van Hell, 2014). Beyond this it was predicted that violation effects would be shown with scalp topographies that were irregular in shape and extent. This ability to assign statistical significance to topographical features represents a novel advantage provided by CForest. These features were predicted to differ between violation types (semantic or phrase structure) and time window (300-500 ms vs. 600-800 ms), resulting in quantification of previously suggested trends (Moreno & Kutas, 2005; Pakulak & Neville, 2010; Weber-Fox et al., 2003). These characterizations were formed for any ID measure that was shown to significantly and substantially modulate the effect of either type of sentence violation. The qualifier of 'substantial effect' was in place because CForest can detect effects which are highly significant, but which may be weak, of little interest, or specific to the data set. Therefore, only ID measures associated with a change in violation effect size beyond 0.5 μV across the observed range of ID measure scores were investigated further. In addition, the added sensitivity associated with bagging and variable pre-selection was expected to reveal relationships that our previous investigations may have been insensitive to. For example, word reading efficiency and speech perception were not previously found to significantly modulate violation effect size, but remained viable candidates for investigation.

Lastly, it was important to test the accuracy of model predictions in novel data on which the model was not trained. As described above, while a model should ideally depict the significance of effects found in data on which it was built, it should also be generalizable to novel data as well. Thus in the present study, models using a range of user-defined specifications were evaluated both in terms of accuracy in a 'learning' data set on which they were built, (randomly selected 80% of our data) as well as generalizability in a 'testing' set (a withheld 20% portion of the data on which it was not built). As discussed, CForest can be sensitive to sample-specific variability, and its propensity for over-fitting is not well-documented. This step was performed to ensure that the resulting model was not solely applicable to describing the data on which it was built. Efficacy in each case was evaluated using the Pearson correlation between the model-predicted responses and those observed, where the ideal set of model specifications was taken as that which performed best under both circumstances. User-defined parameters and their theoretical impact are described below. Briefly, they were the number of trees included in the forest, and the variable selection guidelines in each tree. In each case, there is no defined standard, motivating an exploratory investigation into their impact. It was hypothesized that application of model predictions to novel data would result in a reduced, but significant correlation between predicted and observed responses, speaking in part to generalizability of the model beyond the present data.

## 6.2. Methods

### 6.2.1. Data Acquisition and Pre-Processing

The majority of methodological details are as described in Chapter 2, as the present investigation was a continuation of the prior analysis and therefore used the same data set and pre-processing stages. Unchanged details include a description of the participants included in

these investigations, an overview of the language proficiency and cognition assessments used, the sentence processing task, and EEG acquisition hardware and procedure, and data pre-processing. Therefore, where procedural or analysis details are not described, please refer to Chapter 2. Differences in statistical modeling procedures exist and are outlined in detail below. At times, results will be compared with those reported in previous chapters. Recall of prior results will be outlined where necessary.

### 6.2.2. Conditional Inference Random Forest Analysis

As discussed above, CForest is a collection of techniques which are combined to produce a model of the relationship between a response variable and some number of predictor variables. While the computations for each technique occur in isolation, the results of each form an interdependent (and recursively looping) chain that is the CForest analysis. That is, these processes act in a series of steps which are iterated on increasingly small subsets of a data set to characterize any existing relationships. These processes will be explained in more depth below, but by way of introduction include 1) random selection of data to be subjected to hypothesis testing, 2) determining the predictor variables that are candidates for hypothesis testing in this random selection of data, and 3) recursively partitioning the data into binary subdivisions based on the relationship between the response variable and the variable(s) selected in the second step, provided a relationship exists. These second and third steps are completed in a repeating fashion until no identifiable and significant relationship exists, resulting in a single branching tree structure. The entire process is then completed some user-specified number of times to create a forest. Where a single tree is a simple model which allows for estimation of a response variable using a randomized subset of the full data set, the forest allows for estimation of a distribution of

responses, on which confidence intervals can be developed. Each of the above steps will now be discussed in further detail.

*(1) Bagging (randomized data selection)*. A major drawback of recursive partitioning models, including the approach introduced above (CTree), is their instability to small variations (noise) in the data (Strobl et al., 2009). That is, the hypothesis testing which drives detection of relationships between variables in a single tree relies on a permutation testing framework which is fundamentally different from, and less conservative than, the *F* statistic used in regression models. This framework permutes levels of a predictor variable (either categorical or continuous) to test its association with the response variable, and will be described in more detail below (Strasser & Weber, 1999). The result is that individual trees are susceptible to detecting spurious effects based on outliers. One means to circumvent this problem is to combine CTree modeling with bagging. Bagging is a method by which many random subsets are drawn from the original data set, similar to bootstrapping. These subsets are also known as learning sets. One CTree model (tree) is fit to each learning set, resulting in an ensemble of trees (forest), each based on a unique and randomized portion of the full data set. Bagging also attenuates random effects such as those incurred by repeated measurements and testing site effects (Grandvalet, 2004). For the present study, trees were grown on random subsets each comprising 93.2% of the data. This portion was selected to align with our previous investigation using CForest analysis (McWhinney et al., 2016). Data were sampled without replacement to equalize selection of variables with differing numbers of categories (Strobl, Boulesteix, Zeileis, & Hothorn, 2007).

*(2) Variable Pre-selection*. The second component of CForest analysis is variable pre-selection, the process by which not all predictor variables are candidates for guiding the subdivision of data at a tree node (i.e., branching point in the tree structure, or division of a data partition into two subsets using a predictor's levels as the split criteria). While the mechanics and

theory of the partitioning process will be detailed below, a partition of the data set is subdivided into two smaller subdivisions using levels of a predictor variable as the split criteria (e.g., if age is selected as the partitioning predictor, participants above and below age 40 may be the constituents of the two resulting mutually exclusive partitions). However, the predictor that is selected for guiding subdivision is that which shows the strongest association with the response variable. Therefore, in the presence of one strongly-associated predictor, if all are considered candidates to guide division, then those with weaker (but significant) associations will rarely, if ever, be selected. The result is that only relationships with the most strongly-associated predictors will be described. To alleviate this problem, not all predictors are tested for association. Only a random selection of predictors are evaluated, allowing characterization of the relationship between the response variable and more weakly-associated predictors. This produces more diverse trees, which necessarily improves the generalizability of predictions made at the forest level (Breiman, 2001). The proportion of variables which are randomly selected for inclusion is a user-specified parameter, and a range of proportions were investigated to evaluate this parameter's impact on model quality. The result is that potentially informative interactions that would otherwise be missed can be detected.

*(3) Conditional Inference Regression Trees*: Conditional inference regression trees are non-parametric models that recursively partition data into nodes and branches by way of hypothesis testing. While a single tree allows for an estimated value of a response variable given some number of independent variables, a forest (collection of trees) can be used to generate a distribution of estimates. This distribution can then be used to develop confidence intervals and/or evaluate the significance of contrasts. The CTree algorithm produces a single tree using the following steps (Hothorn et al., 2006):

(1) Global Null Hypothesis (Stop Criterion). Determine whether a subset of data (partition) at the present node should be split into two further subsets by testing the global null hypothesis of independence between observations in that partition and predictor variables. On the first iteration of the branching tree structure, this node includes the entirety of the data that was selected during bagging. As any subsequent nodes only contain the data which has been subdivided through the series of branches that reach it, the association is only tested in that subdivision of the data. As subdivisions necessarily become smaller with additional branches, detecting significant effects becomes less likely in nodes which have been subdivided more times. In addition, the user can specify a minimum number of observations required in a node for hypothesis testing to take place.

The permutation testing framework established by Strasser and Weber (1999) is used to produce the test statistic C, determining the significance of the association between each predictor variable and the response variable. In the event that a significant association exists with at least one predictor variable, the global null hypothesis (that no association exists) is rejected. If the null hypothesis cannot be rejected, subdivision of the data in this branch ceases, resulting in a terminal node.

(2) Variable Selection. If the null hypothesis is rejected in the step above, the predictor with the strongest association to the response variance (i.e., the smallest p-value) is selected for evaluation in terms of how that variable will guide further subdivision of the data, described below. Recall that due to variable pre-selection, not all predictor variables are evaluated. Those which are excluded from consideration in one node may still be evaluated in subsequent (lower) nodes, however.

(3) Best Split Point. The absolute value of the test statistic C is calculated for every possible split point, where this split will be used to subdivide this node's data into two subsequent partitions. For continuous variables, such as age, split points may include any observed age in the data set, where the selected split point is that which maximizes the significance of the response variable contrast between the resulting groups (e.g., the difference in response for participants equal to or above vs. below each observed age). For categorical variables, such as region of residence, all possible comparisons of two groupings of regions are evaluated. For example, one possible contrast might be New York and Boston vs. Halifax, while another might be New York vs. Boston and Halifax. In the present experiment, electrodes were treated categorically and were evaluated in this manner. The data are partitioned at the split point that results in the most significant difference between responses of participants in the two groupings.

(4) Repeat. Repeat steps 1–3 on the two newly created data partitions until either the global null hypothesis cannot be rejected, a user-specified minimum number of observations per division has been reached, or a user-specified maximum number of divisions has been reached.

The process described above is depicted in **Figure 6.1**. The single-tree model results in a series of split points which segregate observations and assign a predicted response in each node. The predicted response is shared by all combinations of predictor levels present at that node. For example, if a tree identifies that participants below age 25 significantly differ in some response from those above or equal to age 25, then it will predict identical responses for any participant under age 25. It is important to note that while divisions are binary, the fact that numerous subdivisions can occur recursively means that several split points can be identified in a single variable. As described, this process estimates a response using a single tree, but each tree is built

using unique combinations of bagged data and randomly selected variables in each subdivision. Therefore a distribution of responses is produced at the forest level. The response variable's value is estimated at every row in the data set using every tree, and then averaged across trees, to produce a summary estimate for each row. This can be compared with observed responses to estimate an $R^2$ value for the model. Finally, the significance of the relationship between these estimates and the predictor variables is determined using the same recursive partitioning algorithm described above, but using the summary estimates instead of observed responses. Any predictor which is associated with a significant split at this stage can be considered to significantly predict the response.

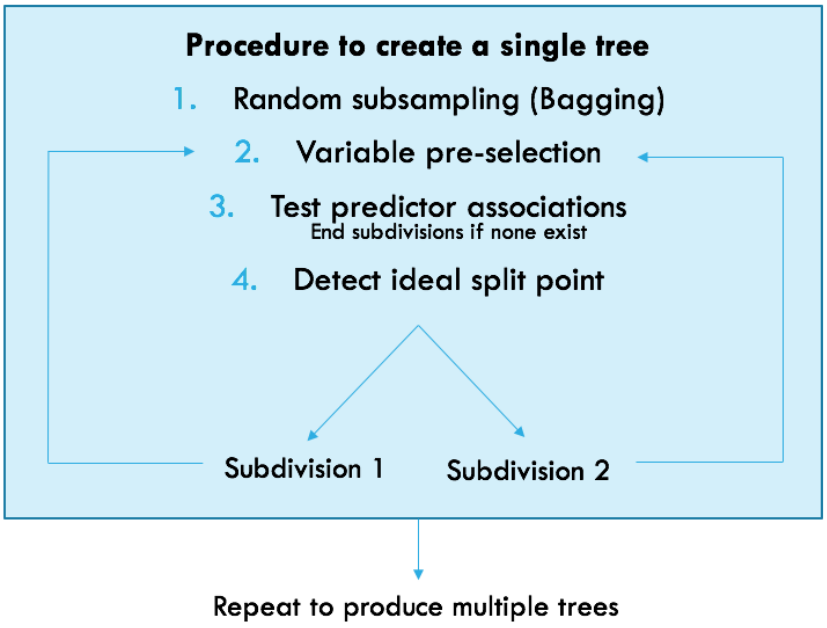


***Figure 6.1*** Flowchart of procedures used to create a single tree. While random subsampling is completed only once per tree, all subsequent steps occur in each of the recursively partitioned data subsets until no significant associations can be detected. This procedure is completed iteratively to for each tree in the forest, where the results in each tree are completely independent of one another.

Using this method, the model Listening/Grammar × Speaking/Grammar × Listening/Vocabulary × TOWRE × AzBio × OSpan × LSpan × Condition (control, violation) × Channel was fitted to the EEG signal amplitude averaged over a time window, for each time window and sentence type (semantic or phrase structure). The input to this model (i.e., dependent variable) was the absolute scalp voltage for each electrode, sentence type (semantic or phrase structure), condition (violation or control), trial, participant, and all ID measure scores for that participant. This format ensured that the two conditions could subsequently be contrasted to characterize the violation effect, and associate its amplitude with any predictors of interest. Predicted responses were produced in correspondence with each individual response value in each condition, so that the two conditions could be contrasted and the predicted responses could be compared with those observed. This model differs from that used in previous chapters by allowing for interactions between the ID measures at the highest-possible level (i.e., potentially 10-way interactions may arise). Previously, each ID measure could only interact with condition and ROI (replaced here by individual channels). This was done to aid in interpretability by limiting the scope of interactions. That is, when using regression, summarizing the effect of a predictor which is involved in complex interactions becomes increasingly cumbersome. However, these potentially high-level interactions were only used to generate the estimated response, and any added nuance in the dependence of the response on these variables would ideally produce a more accurate model. Interpreting the effect of predictors on the response was instead performed only in consideration of that predictor's effects. Moreover, where the estimates of regression models can be invalidated by multicollinearity, CForest is capable of modeling all ID measures (including those correlated with one another) simultaneously as it does not rely on predictor estimates to account for variance. Accordingly, all ID measures were included in a model. This process was completed once in each time window (300-500 ms and 600-800 ms).

The criterion to reject the null hypothesis at each node was set to $\alpha = 0.05$, where probability values were Bonferroni corrected for multiple tests.

### 6.2.3. Optimizing Forest Parameters

Given the multi-faceted nature of the procedures that constitute CForest, a number of aspects of the analytical mechanics must be defined by the user. These include the number of trees in the forest, and the proportion of predictor variables that are considered during variable pre-selection for hypothesis testing in each branch. As this is a novel application of CForest, there is no clear rule on these specifications, and their impact was assessed. We evaluated a range of specifications in each of the three areas to determine their impact on the model, in terms of specificity and generalizability.

In theory, a forest is improved through addition of trees, as this increases the signal to noise ratio and the stability of resulting response estimates for any combination of predictor variables. However, computation time increases with the size of a data set and the number of predictor variables, and so creating forests with excessive trees can be infeasible. We investigated the efficacy of forests built using either 100, 300, or 500 trees. This investigation was conducted to assess whether reducing the number of trees resulted in any loss of forest accuracy or generalizability (to be described below).

In addition to the number of trees, the variable preselection process can impact the efficacy of resulting forests. The number of predictors chosen to evaluate in any node (or more accurately the proportion of total predictors) impacts the likelihood of including predictors with stronger associations in any given tree, and therefore the representation of that predictor across all trees in the forest. Significant predictors with weaker associations can conversely only be described when those stronger predictors are not included for consideration, due to division of

the node only in the predictor with the strongest association. The result is a trade-off between nuanced description of weaker effects (lower number of predictors considered during variable preselection), and accurate portrayal of stronger ones (higher number), where the accuracy of a forest's predictions is maximized at some unknown midpoint. The ideal proportion is specific to a data set, and depends on a predictor's distribution of values and effect size. Therefore, in addition to evaluating the efficacy of forests of varying size, each forest size was also run in consideration of 20%, 40%, 60%, 80% or 100% of predictors during variable preselection.

Both the accuracy of the forest's predictions in the data on which it was built (specificity), and the accuracy in novel data (generalizability), were considerations for each forest. Therefore, only a randomly-selected 80% of our full data set (i.e., response amplitudes randomly selected from any electrode, trial, participant or condition, where each was tied to the ID measures of the associated participant) was used to build the above-described forest. This division was created using a random sampling without replacement of epoched electrode-level measurements across participants, where each measurement coincided with all ID measures associated with a participant. The result is that each of the two data sets very likely contained partial data from all participants. The same random selection was used for all forests to ensure that any variability in quality between models was not due to random differences in the inclusion of outliers in the 'learning' sets. The remaining 20% was set aside to assess the ability of each forest to predict novel responses (the 'testing' set), speaking to the generalizability of the model. In the case of each model built using these data, the Pearson correlation between observed and model-predicted responses was used as a measure of accuracy, and the forest with the highest accuracy in both the learning and testing sets was taken as the best.

Due to the time required to perform these computations, it was not possible to perform these analyses in all four areas of investigation (two time windows by two sentence types).

Instead, we limited our investigation of model parameters (i.e., of the impact of user-specified parameters on model outcomes) to only the semantic violation contrast in the 300-500 ms time window. The ideal parameters were applied to models built for each time window and sentence type contrast (four separate models). Creation of four separate models followed the approach taken in previous chapters.

#### 6.2.4. Summarizing Predictions

When first evaluating one of the four models produced, the presence of a two-way interaction between condition (sentences containing violations or well-formed ones) and region (at the individual electrode level) was investigated using a forest that contained only those two terms. Any electrode showing a significant division of predicted responses between the two conditions was deemed to support the contrast, and the magnitude of the response difference (violation – control) was plotted at those electrodes to qualify the presence of any effects of interest, including the expected N400 and P600 components. Subsequently, a model that also contained terms for ID measures was used to evaluate the influence of these measures on the condition contrast regionally.

The violation effect size was shown in two ways. First, it was investigated as a function of each ID measure globally (i.e., averaged over channels) to identify portions of the ID measure spectrum (similarly scored participants) in which consistent and considerable positivity or negativity was evident at some region of the scalp. These participants were isolated and their average topography for the time window and sentence type was plotted. These two depictions are described below as the *global* and *local* effects, respectively, highlighting that only the latter takes regional constraints into consideration. Distinct groups of electrodes were identified,

providing a grouping of electrodes where this effect was maximal, provided that there was a significant differentiation between electrodes in terms of violation effect size.

Note that this classification allowed a clustering of spatially distant electrodes if similarities existed, owing to the categorical treatment of electrodes. However, this was rarely the case, and spatially contiguous clusters were found in nearly all instances. This method allowed us to identify topographies that were unique in both shape and extent for each predictor, and offered a considerable improvement in sensitivity when defining a region of interest over using pre-defined electrode clusters (e.g., arbitrarily grouping spatially similar electrodes into an arrangement of ROIs). The results presented below demonstrated that the violation effect topography can be irregular in shape or extent, demanding these nuanced methods of detection.

Lastly, the confidence intervals of the forest were shown at the region where the effect was maximal, providing a more localized depiction of the relationship between the ID measure and violation effect size. Confidence intervals were shown alongside the overall predicted response, which showed significant step-wise trends in the underlying variability. Therefore, the overlaid stepping function was the determinant of significant ID measure effects.

## 6.3. Results

### 6.3.1. Forest Specificity

Forests were created using various combinations of two user-defined parameters in order to assess their impact on the accuracy of the model-predicted estimates in terms of the data on which the model was built, or the forest specificity. A separate analysis of the accuracy of the model-predicted estimates in terms of a novel hold-out data set (a random 20% of the initial

data set), or the forest's generalizability, will be described in the following section. As described above, this included variations in the proportion of predictors considered during variable preselection, and the number of trees in the forest. The results of these investigations are described in **Table *6.1***. Quality of the forest's specificity was evaluated using a Pearson correlation between model-predicted estimates and those observed in the 80% of the data set on which these models were built. In all cases, correlations were significant at $p < .001$, (DOF = 6659).

**Table 6.1** Forest accuracy (Pearson correlation of predicted vs. observed responses on which the forest was built) for gradations of variable preselection proportions and tree numbers.

| | | **Number of trees** | | | |
|---|---|---|---|---|---|
| | | ***100*** | ***300*** | ***500*** | *Mean* |
| **Variable pre-selection** | ***20%*** | 0.918 | 0.919 | 0.919 | *0.919* |
| | ***40%*** | 0.919 | 0.919 | 0.919 | *0.919* |
| | ***60%*** | 0.918 | 0.919 | 0.919 | *0.919* |
| | ***80%*** | 0.919 | 0.919 | 0.919 | *0.919* |
| | ***100%*** | 0.919 | 0.919 | 0.919 | *0.919* |
| | *Mean* | *0.919* | *0.919* | *0.919* | |

While increasing the number of trees was intended to improve stability of the estimates, it had no observable effect on the correlation between predicted and observed responses. Increases in the number of trees did coincide with changes in the correlation between predicted and observed responses, but the degree was negligible. A one-way ANOVA revealed that this improvement was not significant ($F(2,12) = 1.49$, $p = .264$). Similarly, while marginal fluctuations in the correlation were noted with changes in the proportion of variables chosen during preselection, these changes were not found to be significant ($F(4,10) = 1.42$, $p = .297$). Therefore,

neither the number of trees used in the forest (between 100 and 50), no the variable pre-selection proportion, were found to influence the specificity of the resulting forest.

### 6.3.2. Forest Generalizability

The procedure detailed above was additionally used to evaluate the accuracy of the model in terms of data on which it was not built, to assess forest generalizability in novel data. This novel 'testing' set contained a randomly-selected 20% of the full data set in order to assess the ability of each forest's output to generalize beyond data on which it was built. Moreover, each forest was built using the same 'learning' set, and tested against the same 'testing' set, to ensure that any variations in generalizability were due to changes in forest parameters. The Pearson correlation was used as a measure of similarity between predicted and observed responses, as when evaluating model fit above. The results of this investigation are outlined in ***Table 6.2***. In all cases, correlations between model-predicted responses and those observed were significant at $p < .001$, (DOF = 6659).

**Table 6.2** Forest generalizability (Pearson correlation of predicted vs. novel observed responses) for gradations of variable preselection proportions and tree numbers.

| | | **Number of trees** | | | |
|---|---|---|---|---|---|
| | | ***100*** | ***300*** | ***500*** | *Mean* |
| **Variable pre-selection** | ***20%*** | 0.677 | 0.678 | 0.678 | *0.678* |
| | ***40%*** | 0.678 | 0.678 | 0.678 | *0.678* |
| | ***60%*** | 0.677 | 0.678 | 0.678 | *0.678* |
| | ***80%*** | 0.678 | 0.678 | 0.678 | *0.679* |
| | ***100%*** | 0.678 | 0.678 | 0.678 | *0.678* |
| | *Mean* | *0.678* | *0.678* | *0.678* | |

As seen in evaluations of model fit above, increasing the number of trees resulted in a modest increase in the mean correlation between predicted and observed responses. However, this increase was not significant ($F(2,12) = 1.71$, $p = .222$). Similarly, there were negligible fluctuations in this correlation across the range of proportions of predictors included during variable preselection, and as above these fluctuations did not show an overall trend, and were not significant ($F(4,10) = 1.699$, $p = .226$). Therefore, neither of these user-defined parameters were found to affect model generalizability.

Given that neither of the two parameters investigated resulted in any significant change to either the model's accuracy in data on which it was built or generalizability to novel data, their specification in the present data was not expected to impact interpretation of results. It is interesting to note that the correlation between CForest's predicted responses in the novel data set and the observed responses for this data set ($r = .6780$), while lower than that for the data set on which the model was built ($r = .919$), was not altogether diminished, and remained significant.

As this evaluation was conducted subsequent to computation of the forests which will be described below, it did not constitute evidence to alter either parameter which had been used. Therefore, all analyses below were completed using 500 trees, at 90% variable pre-selection inclusions. This selection had initially been made to align with our previous use of CForest in neuroimaging data sets of similar size, where variable preselection was allowed to include all predictors except for any one at random (McWhinney et al., 2016). While our prior research was used as a guideline, these parameters were not expected to impact the efficacy of our models.

### 6.3.3. Semantic Violations, 300-500 ms

As described above, a model was first fit to each data set which contained only terms for the condition (those containing a semantic violation or well-formed sentences), and electrode for semantic violations in the 300-500 ms time window. This model revealed that a significant condition contrast was seen across large portions of the scalp. This contrast revealed that sentences containing violations elicited a significantly more negative scalp voltage than well-formed ones across central and parietal regions, while conversely eliciting a more positive response in widespread frontal regions. The latency and distribution of this central-parietal negativity matches that of the expected N400 ERP component, and this interaction is depicted in **Figure 6.2**. It should be noted that for this analysis, region refers to the cluster of electrodes which showed a significant violation effect, as individual electrodes were included as predictors in CForest models. These regions are therefore data-driven in their depiction and should not be confused with ROIs used in previous chapters.

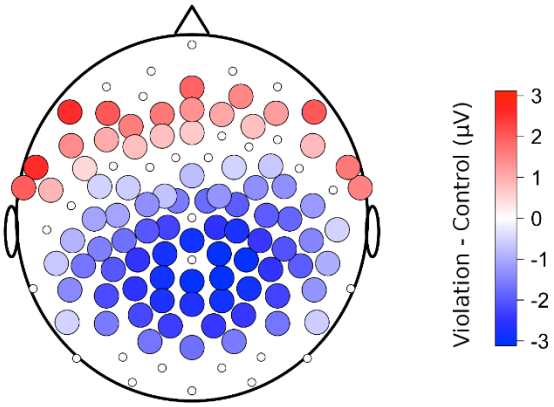


***Figure 6.2*** *Electrodes showing a significant contrast between sentences containing semantic violations and well-formed ones during the 300-500 ms time window are plotted as large circles. Effect size is shown by the colour scale. Smaller electrodes indicate those with no significant contrast.*

Following this, a model was fit which contained terms for condition and electrode as above, but also for each of the ten possible ID measures. This second model was used to evaluate the influence of ID measures on the size of the violation effect, as well as its scalp distribution. Only two of the ten ID measures were found to be related to the size of the violation effect: Speaking/Grammar, and Listening/Vocabulary. The remaining predictors had no impact on the violation effect size for semantic violations in the 300-500 ms time window.

Amplitude of the N400 was greatest in participants with lower scores both for Speaking/Grammar and Listening/Vocabulary. This effect is described in **Figure 6.3** as the *global violation effect*, as it is first presented averaged across electrodes (i.e., globally). The electrodes at which a significant N400 was identified are highlighted as the *peak effect topography*. The effect of these ID measures on N400 amplitude specifically (i.e., without interference from the anterior positivity shown in **Figure 6.2**) was investigated by considering only those electrodes showing a significant N400. The depiction of the ID measures' effect on N400 amplitude in these electrodes specifically will be referred to as the *local violation effect*.

Plotting the local violation effect for Speaking/Grammar and Listening/Vocabulary revealed that a linear relationship between N400 amplitude and ID measure scores was not evident. However, there was a significant trend toward a larger N400 at lower scores. This was depicted as a single division in the stepwise predictions of CForest at approximately the midpoint of each ID measure's scoring spectrum. Note that the global violation effect depicts significant divisions (steps) which were not evident in the local violation effect, though their influence can be seen in the local violation effect's underlying distribution of estimates. Importantly, CForest did not identify these deviations from the identified trend as significant, resulting in a more consistent stepping function with fewer divisions. These fluctuations may have been influenced in part by anterior positivity which was less evident in the local violation effect.

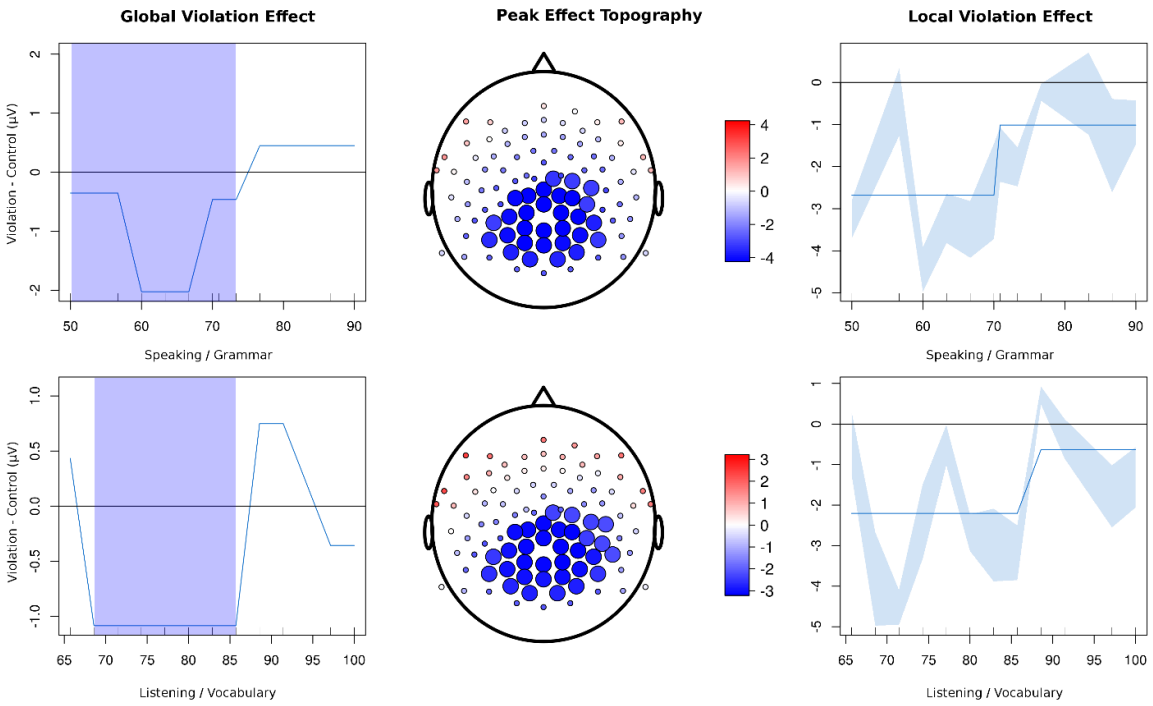


***Figure 6.3*** *Semantic violations in the 300-500 ms time window, showing the violation – control effect (N400) at all electrodes and all participants (left), the scalp topography only for participants who showed the N400 response, as indicated by blue shading in the left pane (middle), and the violation effect in all participants at the electrodes which demonstrated a significant N400 response (right). This series is shown for Speaking/Grammar (top) and Listening/Vocabulary (bottom). All scales are shown in negative (blue) or positive (red) microvolts, where larger electrodes show a significant condition contrast.*

The process for computing the above four plots, including 1) the response topography for all participants, 2) the relationship between an ID measure score and the violation effect across all electrodes, 3) the response topography for a subset of individuals (i.e., those with higher or lower scores), and 4) the relationship between an ID measure score and the violation effect for electrodes showing a significant violation effect as determined in step 1, are outlined in the following pseudo-code. Note that this pseudo-code does not detail the specification of modeling parameters, but provides a procedural outline of the computation of data for each

plot. The structure of this pseudo-code is loosely based on R syntax, with a number of simplifications. In this example, data are plotted for semantic violations in the 300-500 ms time window, and scores are evaluated for a single ID measure. Note that in this pseudo-code, the CForest function refers to the procedure that produces a distribution of estimates equal in volume to the number of trees in the forest, as described in section 6.2 Methods above. The following CTree function refers to the subsequent significance testing of the forest's distribution.

```
# 1) Plot response topography for everyone at all electrodes
plot_topography = function(dat) {
        predicted.raw = CForest(response~electrode*condition, data=dat)
        predicted.steps = CTree(predicted.raw~electrode*condition)
        predicted.steps.v = predicted.steps[condition='violation']
        predicted.steps.c = predicted.steps[condition='control']
        electrodes.sig = electrodes[predicted.steps.v!=predicted.steps.c]
        plot_head(add_color=electrodes.significant)
}

dat = dat.semantic.300_500
plot_topography(dat)

# 2) Plot global violation effect (relationship across all electrodes)
plot_effect = function(dat) {
        predicted = CForest(response~score*condition, data=dat)
        predicted.v = predicted[condition='violation']
        predicted.c = predicted[condition='control']
        predicted.v-c.raw = predicted.v - predicted.c
        predicted.v-c.steps = CTree(predicted.v-c)
        plot(score, predicted.v-c.raw)
        plot(score, predicted.v-c.steps)
}

dat = dat.semantic.300_500
plot_effect(dat)
```

```
# 3) Plot response topography for select individuals
dat = dat.semantic.300_500[score>60]
plot_topography(dat)

# 4) Plot local violation effect (relationship at sig. electrodes)
dat = dat.semantic.300_500[electrode=electrodes.sig]
plot_effect(dat)
```

### 6.3.4. Semantic Violations, 600-800 ms

A model that contained only terms for the condition contrast and electrode showed that, for semantic violations in the 600-800 ms time window, a significant positive condition contrast was found to be widespread and primarily central in topography. Anterior left and right scalp regions also demonstrated significant negativity in the condition contrast. This effect is outlined in **Figure 6.4**.

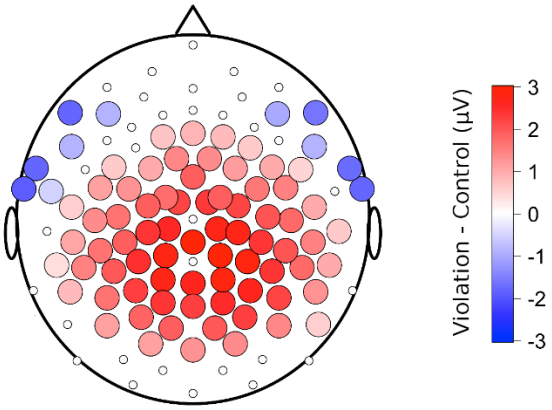


***Figure 6.4*** *Electrodes showing a significant contrast between sentences containing semantic violations and well-formed ones during the 600-800 ms time window. Smaller electrodes indicate those with no significant contrast.*

Second, a model which contained both of the above terms (condition and electrode) as well as each of the ten potential ID measures was fit. However, none were found to influence either the amplitude of this response or its topographical distribution.

### 6.3.5. Phrase structure violations, 300-500 ms

A model that contained only terms for the condition contrast and electrode for phrase structure violations in the 300-500 ms time window showed a small number of electrodes at which a significant positive or negative violation effect was seen in response to phrase structure violations during the 300-500 ms time window. A modestly-sized cluster of electrodes in the left parietal region demonstrated a significant negative violation effect, with a more diffuse cluster showing a positive violation effect in right temporal and parietal lobe regions. This pattern is shown in **Figure 6.5**.

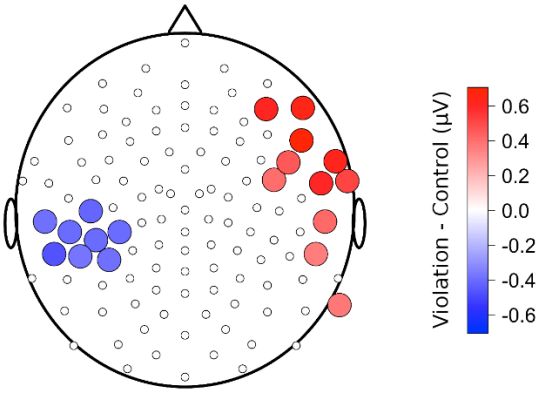


***Figure 6.5*** *Electrodes showing a significant contrast between sentences containing phrase structure violations and well-formed ones during the 300-500 ms time window. Smaller electrodes indicate those with no significant contrast.*

Second, a model which contained the terms described above, but also each of the ten ID measures was fit. Phrase structure violation effects in the 300-500 ms time window were

influenced by Listening/Grammar and Listening/Vocabulary. Other ID measures were not associated with changes in the violation effect size. As with semantic violations, these effects were considered first in terms of the global violation effect (the interaction between ID measures and violation effect size, averaged across electrodes), second in terms of the topography for individuals who demonstrated a significant positive violation effect (peak effect topography), and third in terms of the relationship between ID measures and violation effect size for all participants, specifically in the electrodes which demonstrated a significant positive violation effect (local violation effect). These relationships are each detailed in **Figure 6.6.**

Regarding the global violation effect, the lower half of Listening/Grammar scores were associated with a positive violation effect. Those individuals highlighted as being representative of this response demonstrated a violation effect topography covering an arc-like distribution of electrodes, sweeping from anterior midline to left parietal regions of the scalp. The local violation effect for these electrodes showed a positive violation effect, where the underlying distribution of estimates depicted an intermittent peak that was associated with a single participant and was not deemed significant.

A similar pattern to the above was seen in the relationship between Listening/Vocabulary and violation effect size, with the exception that this significant positivity was only found in the three participants with the lowest Listening/Vocabulary scores, as seen in the global violation effect. The topography of this effect for these participants was widespread, with the response being strongest in left anterior electrodes. Given that the local violation effect for this ID measure contained nearly all electrodes, it closely matches the function depicted in the global violation effect.

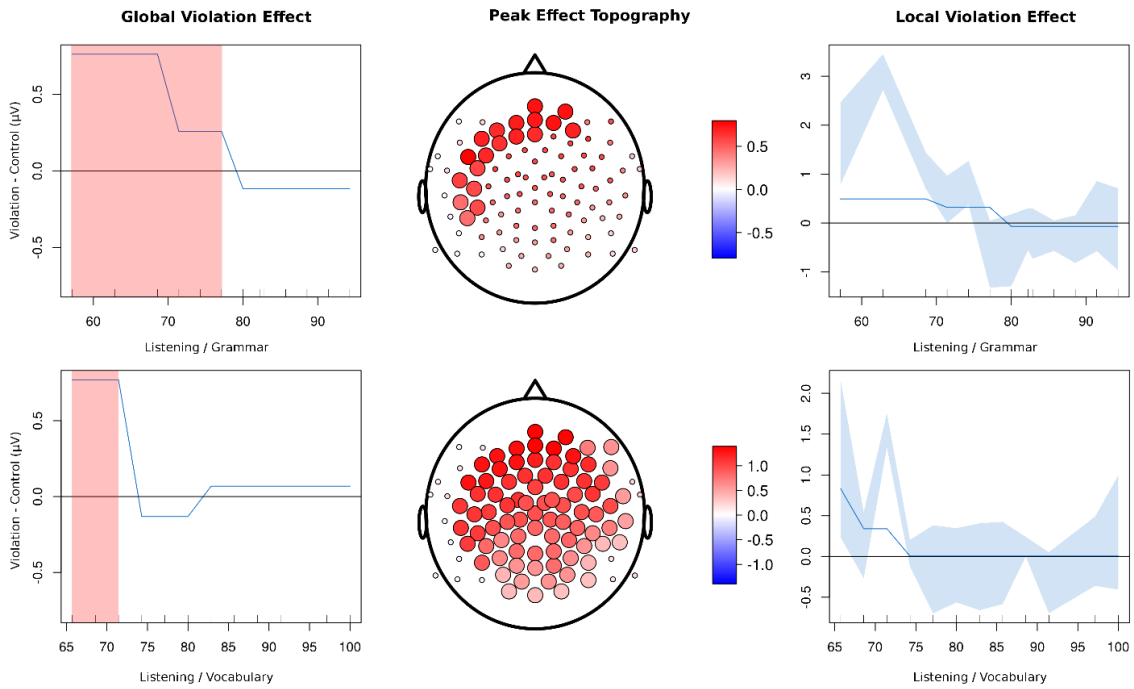


***Figure 6.6*** *Phrase structure violations in the 300-500 ms time window, showing the violation – control effect at all electrodes and all participants (left), the scalp topography only for participants who showed the response (middle; those shaded in the left pane), and the violation effect in all participants at the electrodes which demonstrated a significant response (right). This series is shown for Listening/Grammar (top) and Listening/Vocabulary (bottom). All scales are shown in negative (blue) or positive (red) microvolts, where larger electrodes show a significant condition contrast.*

### 6.3.6. Phrase structure violations, 600-800 ms

A model was first fit which contained only terms for condition and electrode for phrase structure violations in the 600-800 ms time window. In this model, the two-way interaction between these terms revealed that phrase structure violations elicited a generally positive response in the 600-800 ms time window. The distribution of this response was consistent with the expected P600. This effect was widespread and primarily central in topography. In addition, a

small cluster of left anterior electrodes showed a significant negative violation effect, along with a single electrode in the right central scalp region. This pattern is shown in **Figure 6.7**.

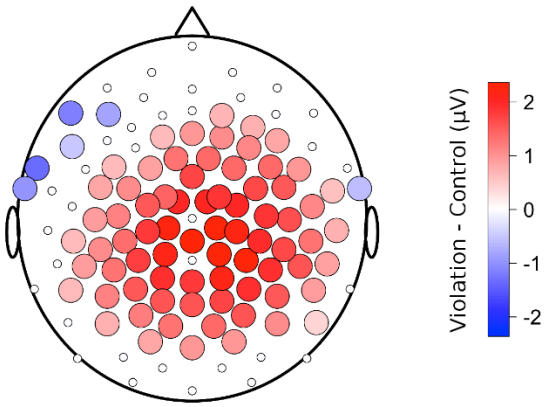


***Figure 6.7*** *Electrodes showing a significant contrast between sentences containing phrase structure violations and well-formed ones during the 600-800 ms time window. Smaller electrodes indicate those with no significant contrast.*

Second, a model was fit using the terms described above, in addition to a term for each of the ten ID measures. However, TOWRE was the only ID measure that was found to influence P600 amplitude, as there was an association between higher TOWRE scores and increasing P600 amplitude. This can be seen in the global violation effect depicted in **Figure 6.8**. The topography of the violation effect for those individuals who showed a P600 response revealed that there was a significant condition contrast, which was strongest at central midline electrodes and far-reaching to lateral parietal regions. The local violation effect revealed that while there was considerable variability in P600 amplitude across participants, the trend toward stronger positivity at higher TOWRE scores resulted in a single significant division between lower- and higher-scoring participants, the latter of whom showed stronger P600 amplitude.

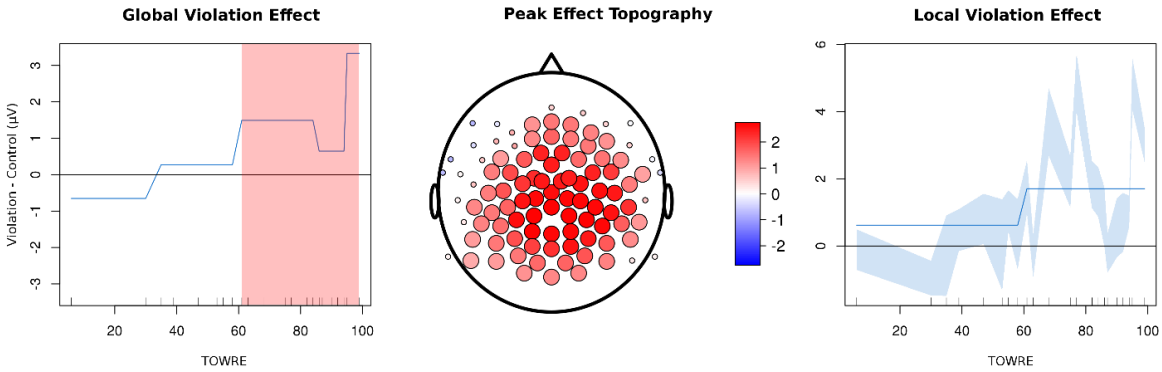


***Figure 6.8*** *Phrase structure violations in the 600-800 ms time window, showing the violation – control effect at all electrodes and all participants (left), the scalp topography only for participants who showed the response (middle; those shaded in the left pane), and the violation effect in all participants at the electrodes which demonstrated a significant response (right). This series is shown for TOWRE, as it was the only ID measure found to influence P600 amplitude. All scales are shown in negative (blue) or positive (red) microvolts, where larger electrodes show a significant condition contrast.*

## 6.4. Discussion

### 6.4.1. Overview of Objectives

The present study aimed to improve characterization of how ID measures affect ERP components associated with various aspects of language processing, in several ways. First, through removing user subjectivity when determining proficiency 'bins' (i.e., proficiency windows in which a sentence violation effect is seen). Second, through data-driven detection of ROIs that are potentially irregular in shape and extent, and the shape/extent of which can be deemed statistically significant. This step provided quantification of previously-qualitative evaluations of the topographical distributions of effects (e.g., Pakulak and Neville, 2010). Third, in providing an approach that is more robust to multicollinearity, irregularities in distribution, sample size or residuals, all of which are problematic to regression approaches (particularly in

over-fitting to sparse sample distributions, as seen using GAMM). We have investigated the ability of CForest to accomplish each of the above needs, and will evaluate its strengths as well as considerations for, and limitations on, its use.

### 6.4.2. Summary of Semantic Violation Effects

First, it was necessary to identify two components of interest (N400 and P600) in order to investigate CForest's characterization of their dependence on ID measures. Each type of violation (semantic and phrase structure) was assessed in two time windows: 300-500 ms, and 600-800 ms, following the onset of the violation. While the effect of semantic violations was investigated in both time windows, this type of violation was primarily intended to elicit an N400 response, consisting of a central-parietal negativity during the 300-500 ms time window.

CForest depicted a significant contrast between sentences containing semantic violations and well-formed ones during the 300-500 ms time window, showing a relative negativity associated with sentences containing violations. The topography of this effect was consistent with the expected N400 response. The amplitude of this response was further found to be predicted by Speaking/Grammar and Listening/Vocabulary scores on the TOAL-3 language proficiency test, with lower-scoring individuals showing a larger-amplitude response. This finding was largely consistent with those suggested using LME, except in the LME analysis the directionality of this interaction was reversed (i.e., higher-scoring individuals showed a higher amplitude response for each of Speaking/Grammar and Listening/Grammar). The results of our GAMM analysis, however, did not suggest a response that was consistent enough across regions or ID measures to interpret, likely due to sample size concerns as described in the previous chapter. In addition, the present findings were consistent with some previous work (Moreno & Kutas, 2005; Weber-Fox et al., 2003). However, findings in this area have been inconsistent, as a

stronger N400 response has also been associated with higher-proficiency individuals, rather than lower, both in native English speakers and those who learned English as a second language (Newman et al., 2012). These inconsistencies highlight the need for further research in this area.

Similar to the results of our earlier LME analysis, CForest revealed that semantic violations were also associated with a significant positive violation effect during the 600-800 ms time window. However, the amplitude and distribution of this response were not found to be associated with any of the ID measures investigated. While LME found three ID measures to significantly influence semantic violation effect amplitude during this time window, Speaking/Grammar scores showed the strongest influence. We reported similar findings for our GAMM model of these data. In this case, CForest may have been insensitive to the gradual slope of the influence of Speaking/Grammar scores on response amplitude.

#### 6.4.3. Summary of Phrase Structure Violation Effects

In addition to the above, the effect of phrase structure violations was investigated in both time windows. Although we posed no hypotheses regarding the effect of ID measures on the response during the 300-500 ms time window, the amplitude of this response was largest (most positive) in participants with lower Listening/Grammar and Listening/Vocabulary TOAL-3 scores. This finding mirrors our previous GAMM results, which suggested a stronger response in individuals with lower Listening/Grammar scores, particularly in left/anterior regions. LME suggested a similar finding for individuals with lower Listening/Grammar and Listening/Vocabulary scores, but conversely that the response was instead strongest participant with higher Speaking/Grammar scores.

In the 600-800 ms time window, a positive violation effect was obtained, with a topography consistent with that of the expected P600 response. The only ID measure that was

found to influence P600 amplitude was TOWRE, where the P600 effect was found to have the largest amplitude in higher-scoring individuals. This finding was divergent from that of the other analysis methods used, which predicted significant influence of several ID measures. While the patterns described by LME and GAMM in this instance were remarkably similar to one another, neither found TOWRE to be a significant predictor of P600 amplitude. Importantly, due to the variable selection process that was used for both approaches, TOWRE had been excluded due to high collinearity with OSpan, which was considered the better predictor in terms of model likelihood.

#### 6.4.4. Optimizing Forest Parameters

It was important to quantify the effect of any user-defined parameters on the resulting forest, in order to avoid subjectivity where possible. Aside from the model terms, which have been rigorously defined and justified both here and in previous chapters, two parameters had the potential to impact model fit and generalizability. These were the number of trees included in the forest, and the proportion of predictors included during variable preselection. Importantly, changing either parameter had no appreciable impact on the model in this case. The effect of the number of trees computed is similar to that of the number of samples prepared during a bootstrapping procedure, as each represents a randomized subset of the data on which estimates are built. Increasing the number of randomized subsets should be expected to improve model accuracy, especially in smaller sample sizes where variability between subsamples can be expected to be higher. Given the size of the present data set (nearly 1.5 million observations), it is perhaps not surprising that varying the number of trees did not significantly impact either the accuracy of model fit or generalizability in novel data. These findings suggest that as few as 100

trees may be adequate in a typical EEG data set such as the one used in the present study, though it is difficult to generalize to smaller datasets.

A considerably less straight-forward parameter is the variable preselection proportion. Depending on the size of the data set, the number of predictors, their distribution, and their association with responses, changes in this parameter can impact model outcomes (Breiman, 2001; Hothorn et al., 2006). Limiting trees to inclusion of only a few predictors ensures that, if a small number of predictors are considerably more strongly associated with the response than all others, more trees will be generated that do not include those dominant predictors. This allows for significant divisions to be identified in predictors that have weaker overall influence on the variances of the data. Detection of weaker effects is therefore only possible at a degree of expense to accurate portrayal of stronger ones, and the ideal proportion balances the two in a way that is specific to a data set. Not only can this impact characterization of results, but also accuracy of the model, and by extension generalizability in novel data. Given the nature of this trade-off, a range of proportions should ideally be investigated in any analysis using CForest. Despite the theoretical importance of selecting the appropriate proportion, a wide range of specifications were evaluated and none were found to significantly impact our model's efficacy. Therefore, while no standard exists for this proportion or the number of trees, neither was considered salient in the present analysis.

#### 6.4.5. Interpreting Significance with CForest

When using CForest to describe variables such as those used in the present study (electrodes, conditions, or continuous variables), it is important to consider that the output depicts significant differences in a response based on divisions in a predictor. The result is either the significance of a conditional contrast for categorical variables, or piecewise-constant

segments in continuous variables. By comparison, additive modeling solutions such as LME or GAMM use either a linear slope or a smooth fit formed by a series of polynomial basis functions to describe the relationship between a continuous predictor and response variable. Therefore, despite variability in the forest estimates in the violation effect which may or may not be significant, any changes which are depicted in CForest's final stepping function across an ID measure are considered significant deviations. This provided clear-cut proficiency bins, which were used to plot response topography. While this removes a degree of subjectivity that would be required to define "high" or "low" proficiency bins, which has been done heuristically or *a priori* in the past (e.g., Pakulak and Neville, 2010), this characteristic of CForest also represents a limitation. Underlying nonlinearities in the response amplitude can only be described through inconsistencies in the stepping function produced by CForest, as a smooth fit cannot be produced. Likewise, even linear trends must be depicted as a series of steps of consistent height. While this has been useful in CForest applications aimed toward categorical determinations, and here was used to define cut-offs for proficiency bins or ROIs, this approach may be less suitable for describing continuous variables.

Using the present approach, nonlinearities that underlie CForest's stepping functions may be ascertained using the confidence intervals of the forest, which were here plotted for the local violation effect in any interaction. However, caution must be taken in interpreting these predicted confidence intervals. Notably, the underlying function is strongly influenced by individual variability, particularly in samples with only a single observation unit (participant) for any given predictor value, and even large fluctuations may not be significant. This was frequently the case in the present study, as participants rarely scored identically on proficiency measures, resulting in confidence intervals often based on a single participant, which were accordingly wide. Therefore, the underlying function may be strongly representative of the sample, but not

generalizable. This motivated assessment of the overall trend though the permutation test that is depicted alongside the confidence intervals (i.e., the stepping function showed alongside the forest's confidence intervals), which indicates the significance of changes in the underlying predicted amplitude.

### 6.4.6. Depicting Topographies

In addition to data-driven proficiency bin characterization, CForest allowed for depiction of effect topographies with finer detail than has been possible in previous analyses. Specifically, regions of interest did not need to be defined *a priori*, and could include any possible combination of electrodes, regardless of spatial proximity. Grouping of electrodes was performed entirely through similarity of response patterns. Considering the number of possible groupings that 128 channels could produce, this level of sensitivity in topographical plotting is simply not possible using regression approaches, which would require inordinate corrections of significance for multiple comparisons. Even the mass univariate approach of correcting for 128 individually-assessed electrodes would result in a considerable loss of sensitivity (Groppe, Urbach, & Kutas, 2011). The result is that the significance of even subtle shifts in shape, extent or laterality of effects can be described, where this has not been possible in the past. This advantage would be of value, for example, where Pakulak and Neville (2010) described the P600 to be more broadly distributed in higher-proficiency individuals, but were only able to describe the effect qualitatively, in the absence of a method able to address the question. In the present data, our results did not motivate investigation of a low-proficiency P600 topography in response to phrase structure violations, as only higher-proficiency individuals demonstrated a violation effect with significant positivity. Nonetheless, this remains a compelling advantage of CForest.

Concerning ROIs, a fundamental distinction should be drawn between those identified using CForest and their common use in ERP research. That is, while a region is typically determined *a priori* to be of interest in, the move to data-driven topography depiction results in a model-defined ROI when using CForest. Moreover, there are several important considerations in interpretation of ROIs depicted by CForest. First, as described, there is no requirement of spatial proximity or distribution of any kind in electrodes that are considered an ROI. Clustering of electrodes is performed purely in analysis of correlations in activation magnitude and variability, treating individual electrodes as nominal and categorical factor levels. Therefore, it is entirely possible that two distant groups of electrodes be grouped, which may indeed be expected in some circumstances. However, the fact that violation effects were consistently found to cluster in spatial proximity and follow expected distributions is in keeping with expectations for the present study. The interaction of Listening/Grammar and phrase structure violation effect size during the 300-500 ms time window represents an interesting case, where a positive violation effect was seen across the majority of the scalp, with a central and parietal distribution. However, when evaluating the topography for lower-scoring participants, only the N400 response was only found to be significant in the left/anterior electrodes. This grouping is likely rooted in lower variability of the response at left/anterior electrodes than elsewhere. This case demonstrates another advantage of data-driven ROI selection, in that such a grouping may be meaningful, but is unlikely to be selected *a priori,* or to be obvious when only looking at topographical averages.

### 6.4.7. Ideal Use Cases for CForest

We previously suggested that the sinusoidal functions occasionally depicted using GAMM resulted from over-fitting to subject-specific variability. CForest depicted a similarly

sinusoidal dependence of violation effect amplitude on some ID measures, but this variability was rarely found to be significant. In this regard, CForest may represent a means of overcoming problems associated with over-fitting that GAMM is more susceptible to, as the influence of individual variability on response estimates is balanced with the requirement of consistency when identifying significant subdivisions using a continuous variable. That is, while the confidence intervals of the forest showed considerable noise in some effects (i.e., sinusoidal functions), CForest did not deem this noise significant, resulting in either a single step following the overall trend of the function, or none at all. GAMM, however, was unable to provide such a determination. The present approach may therefore be a capable method of non-parametric effect characterization in sample sizes that are problematic for GAMM.

Given these considerations, CForest may be less susceptible to over-fitting than GAMM. However, as discussed above, the cost is a more blunt characterization of continuous variables. Indeed, while the stepwise and categorical divisions that CForest provides can be invaluable for determining cutoffs where detailed post-hoc testing would not be possible using regression (e.g., depicting the significance of an effect's topography at the level of individual electrodes), this very characteristic may be a detriment to assessment of continuous variables. Future research may therefore use a combination of approaches, defining scalp topography or sentence type contrasts using CForest and then further investigating the effect of ID measures using regression approaches such as LME or GAMM. The stepwise segmentation of an ID measure's effect on a component's amplitude may be desirable in some cases, such as defining bins of high or low scores on ID measures. However, we feel that the more detailed description of continuous variables that regression can provide will likely be better suited to the typical ERP study.

### 6.4.8. Sample-Dependent Effects

A critical consideration during interpretation of CForest results is that, as described, the significance of an effect is not determined in relation to zero microvolts, but instead in relative differences between observations. When comparing two conditions (control vs. violation), detecting a significant difference is comparable to finding a significant condition effect using regression. However, this same hypothesis testing mechanic was used to identify which electrodes showed the violation effect most strongly, out of a considerably more complex set of possible electrode permutations. In both cases, the determination of significance was based on unique properties of the variance in the response amplitude, either between conditions or between sets of electrodes. In the case of electrodes, results can be misleading if the reader interprets the effect's topography as indicating those electrodes which showed a significant violation effect (i.e., a P600 effect with amplitude significantly greater than 0 μV). Rather, this topography was used to show where the violation effect was significantly higher than *other electrodes outside that region*, in the same way that the control and violation conditions were compared. For this reason, variability in response amplitude across the scalp can influence the shape and extent of the effect's topography, and to the extent that this variability is dissimilar between sentence types or time windows, comparisons of topography between the two cannot be made.

While CForest is generally considered robust to heteroscedasticity (Hothorn et al., 2006), differences in the variability of a response either between conditions or in high vs. low-scoring individuals may still influence the significance of effects for the reasons described above. With this in mind, any comparisons should be interpreted with caution. Susceptibility to these issues is widespread in statistical modeling techniques, and so regardless of approach, attention must be paid to randomization of experimental conditions where possible.

### 6.4.9. Model Generalizability

Despite an expected reduction in the correlation between predicted and novel observed responses when compared with observed responses from the 'learning' set, this correlation remained highly significant. This finding lends support to the generalizability of the model, but should be interpreted in mind of two important caveats. First, predictions can only be made for predictor values (ID measure scores) on which the model was trained, although these can be derived from novel combinations. For example, the model could be trained on two participants, one of whom performed poorly on two tasks and one of whom performed well on both. Model predictions could then be applied to a third participant who scored poorly on one and well on the other. However, it is not possible to predict responses for participants with ID measure scores that the forest had never encountered. Considering our relatively limited participant pool (and generally non-overlapping distribution of ID measure scores), this meant that the 'testing' set could not include entirely novel participants, but instead was composed of a randomized selection of responses from trials and/or electrodes in participants who were also present in the 'learning' set.

While not every instance of a continuous variable is required to produce a predictive model, in this instance, removal of a participant entirely from the 'learning' set would have resulted in a considerable gap in ID measure scores due to our limited number of participants. This was avoided as a wide array of missing values might make the model untenable. Importantly, however, there is a distinction between regression and CForest in how previously-unseen scores would be evaluated when estimating a response. Using LME, a positive linear estimate might be extrapolated beyond the highest score that the model had evaluated to estimate a stronger response. Conversely, using CForest, an ID measure score beyond the bounds of what the 'learning' set had been fit to would be assigned an estimated response equal to the

highest-seen score. CForest may therefore be entirely inappropriate for extrapolating beyond the bounds of the predictors on which it was trained.

Importantly, despite the fact that data from some participants were divided amongst the two data sets, the data sets were entirely distinct in terms of individual observations. Ideally, participants would be entirely unique across the two. However, accurate prediction of responses in novel trials and scalp regions, even in shared participants, still speaks to generalizability of the forest. This issue could only be overcome with a much larger sample size, or through considerable down-sampling of ID measure scores (i.e., translation from raw scores into high vs. low performers). The reduction in variability that would result from the latter would represent a substantial loss of information, however. The second noteworthy caveat is that no clear cut-off exists in evaluating the strength of a correlation when deeming a model 'accurate'. In all cases, however, that seen between the predicted and novel observed responses was significant, suggesting efficacy of the present model, and promoting future investigation in larger samples.

#### 6.4.10. Conclusions

In conclusion, we have demonstrated a number of advantages that can be achieved through characterizing proficiency-related language violation processing using CForest. Data-driven ROI selection has proven capable of detecting the shape and extent of violation effect topographies, while largely circumventing concerns surrounding heteroscedasticity. However, in many cases the piecewise-constant predictions that CForest provides are not ideal for depicting nuanced effects of continuous variables. CForest appeared insensitive to gradual linear slopes of the influence of ID measure scores on violation effect amplitude in a number of interactions, which both LME and GAMM were able to detect adequately characterize. However, this same binary partitioning mechanic which is not strongly suited to describing linear influences can be

advantageous in detecting clear-cut effect ranges (i.e., grouping low- or high-scoring individuals) without problems associated with over-fitting, which have otherwise proven problematic for GAMM.

# Chapter 7: Discussion

## 7.1. Overview of Research Objectives

The present research explored a variety of approaches to characterizing individual differences in language processing. At present, while individual differences in language proficiency and working memory capacity have been associated with the latency and topography of the N400 and P600, results have varied between studies. For example, while it has been suggested that individuals with lower scores in language proficiency assessments demonstrate a stronger N400 response to violations of semantic expectations during sentence reading tasks (Moreno & Kutas, 2005; Weber-Fox et al., 2003), precisely the opposite relationship has also been demonstrated, with higher-proficiency individuals showing a stronger response (Newman et al., 2012). These studies provide convincing evidence that individual differences do play a role in N400 production during sentence reading, but inconsistencies in findings have motivated an investigation into methodological considerations surrounding modeling approaches.

The approaches taken in previous research have primarily used either ANOVA (Moreno & Kutas, 2005; Pakulak & Neville, 2010) or fixed-effect multivariate regression models (Tanner et al., 2014; Tanner & Van Hell, 2014). Only one study to our knowledge has used linear mixed modeling (Newman et al., 2012). However, these techniques present a number of constraints in terms of modeling capabilities which may impede interpretation of results. The methods used in this dissertation were selected to address these limitations in a number of ways, and while none are ideal for all purposes, each presented a unique means to overcome some of the concerns in

these areas. While the findings of each analysis will be detailed in greater depth below, their general aims were as follows.

First, we tested a method which is mechanically similar to that used in fixed-effect multivariate regression (Tanner et al., 2014; Tanner & Van Hell, 2014), but with support for modeling sources of random variability – linear mixed effects modeling. This method has been used successfully to describe IDs in language processing in the past (Newman et al., 2012), and served as a baseline analysis to compare subsequent methods with. Using LME, we investigated the model selection process, both in terms of selecting a set of ID measures that are comprehensive but maximally orthogonal, and also in terms of specifying a random effect structure that is appropriate for the experimental design and analytical requirements. Whether the expected N400 and P600 had been elicited, and the influence of ID measures on their characteristics, were also explored in relation to the literature which motivated these investigations.

Second, we used a technique which expands on the capabilities of LME by relaxing the assumption of linearity in the relationships between these measures and response amplitude. For this analysis, we used generalized additive mixed modeling. Rather than fitting linear predictors to a set of observations, this method fits series of smooth polynomial functions that demonstrate a user-specified degree of flexibility. While it was predicted that allowing for nonlinearities in these associations would reveal important details regarding the role of IDs in N400 and P600 presentation, findings resulted in concerns that the present sample size was not sufficient for such a fine-grained description. This analysis was therefore presented alongside a simulation of additional data to characterize the influence of sample size on the amplitude and smoothness of nonlinear functions.

Lastly, we investigated whether moving to a method of data-driven effect characterization would allow for a more nuanced and informative analysis of response amplitude or topography in relation to IDs. For this analysis we used conditional inference random forest modeling, or CForest. This is a nonparametric permutation testing approach which allows for more flexibility in identifying interactions than regression models would allow by relying on a unique framework for hypothesis testing. In addition, an assumption of linearity in relationships is not required. Appropriate use cases and limitations of this approach are discussed, with advice for implementation in future studies. In the following pages we will evaluate the results of each approach and their implications in greater detail. Following this, the overall findings of the effect of IDs will be discussed in relation to the literature, with considerations for future research.

## 7.2. Linear Mixed Modeling

Using LME, we first explored the model selection process. Notably, this process is not specific to LME, and a similar framework might be used with models that do not include random effects (i.e., linear fixed-effect only models). As these analyses included a variety of assessments of language proficiency and other cognitive functions, it was necessary to address multicollinearity between measures and develop a rigorous approach to including a set of measures which balances a comprehensive overview of individual characteristics with choosing a set that is maximally-orthogonal. As this was a largely exploratory investigation, more measures were included than might have been necessary, complicating the model selection process. Future research which consistently identifies specific measures as unimportant — or hypothesis-driven studies that focus on a single or limited range of measures — may help to streamline the selection of measures to be included in an analysis.

These goals of reducing multicollinearity and maximizing a model's descriptive ability represent a trade-off, as reducing the number of measures improves the former at the detriment of the latter. Moreover, there is no standard approach to selecting the ideal set of measures. This step is critical in any ERP study, in order to identify measures which may conflict in describing variance in the data. While collinear predictor variables do not hurt the predictive ability of a model overall, it can result in errors of effect size estimation for individual predictors, which was indeed one aim of this dissertation (Bollinger et al., 1981). In the present data, the results identified three pairs of related measures, each requiring elimination of one: (1) Listening/Vocabulary vs. AzBio (speech comprehension); (2) TOWRE (word reading efficiency) vs. OSpan (visual working memory), and (3) Speaking/Grammar vs. LSpan (auditory working memory). Note that this pairing is based on a modest sample size, and may differ in future studies. Moreover, while the similarity of any two measures may be meaningful, the interpretation of such correlations (i.e., mechanisms driving the relatedness of word reading efficiency and working memory) is beyond the scope of this research. From a modeling perspective, however, elimination of one measure from each pair should aid the accuracy of effect estimation for remaining predictors (Bollinger et al., 1981).

The decision not to allow ID measures to interact with one another in models was made primarily to aid interpretability. That is, we were interested in differences in response amplitude over the scalp between higher- and lower-scoring individuals on any given ID measure, but not necessarily in the interactions between ID measures. Aside from difficulties in interpreting the meaning of such four-way (or higher) interactions (e.g., violation × region × OSpan × AzBio), the challenges associated with post-hoc testing were considered too great in spite of any additional merit these interactions would provide. Specifically, we were concerned that this level of post-hoc testing would require a degree of multiple comparison correction that would diminish the

sensitivity of our analysis to the interactions of interest. While our later investigations using CForest would be less hampered by these issues for reasons discussed in Chapter 5.4, the decision to disallow any interaction between ID measures in that analysis was made to maximize our ability to compare findings between the three approaches. Despite this, research with predictive (rather than descriptive) aims, for example predicting response characteristics or ID measure scores rather than describing relationships between the two, might benefit from modeling such high-level interactions.

The random effect structure that was identified using LME was chosen with the suitability for the experimental and model design in mind. It was decided that creating separate models for each time window and sentence violation type was suitable in order to aid with interpretability and reduce the degree of post-hoc testing required by removing sentence type and time window as dimensions. Therefore, the random effect structure was required to match this decision. That is, while individuals might be expected to produce random variance on these dimensions, modeling such sources as random effects was neither required nor possible. Future research which might incorporate multiple time windows or sentence violation types into a single model, however, may find that incorporating random effects related to either of these two factors improves the model's likelihood. Indeed, it might be expected any model which includes interactions between all of these effects would likely benefit from a maximally descriptive random effect structure (Barr et al., 2013). Such a model would also allow for significance testing on these removed dimensions, which was not central to our aims.

In our analysis, it was demonstrated that model likelihood was considerably improved through allowing for a random variation in response amplitude between participants. This finding was unsurprising, given that allowing for a degree of flexibility in this regard should move some variance from the error term to this random effect, as it could not be accounted for

through the fixed effects of the model. Furthermore, allowing predicted response amplitude variance across the scalp to differ between participants resulted in an additional improvement to model likelihood. This analysis demonstrated that while accounting for random effects through the use of mixed models is not currently the standard in ERP research investigating individual differences – and indeed no standard may yet exist – doing so may yield improved sensitivity to subtle effects.

## 7.3. Generalized Additive Mixed Modeling

Our analysis using LME was followed with an exploration of the assumption of linearity in this area using generalized additive mixed modeling, or GAMM. Specifically, we aimed to reveal whether relaxing the assumption of linearity in the association between ID measures and response amplitude would yield improved sensitivity to subtle effects, allowing us to identify important details regarding the nature of individual differences in violation processing. Notably, at the time of processing, we were unable to directly compare the significance of the violation effect (the difference between the violation and control smooth terms) as this functionality is not native to the package used for modeling. It has since become apparent that this type of computation might be achieved using the *itsadug* R package (van Rij, Wieling, Bayyen, & van Rijn, 2017). However, given the concerns surrounding over-fitting in this analysis, assigning *p* values to these difference curves was not expected to aid in interpretation of findings.

While model fit demonstrated nonlinearities in these relationships, and at times appeared to conform to expectations, the influence of several measures on response amplitude was depicted as sinusoidal. Importantly, no prior research to our knowledge has demonstrated this type of effect of individual differences on the N400 or P600. Therefore, while there was evidence that suggested the influence of IDs on response amplitude might be nonlinear, the

validity of these findings was called into question by fits which appeared overly-specific to our sample. This possibility raised concerns that our sample size was not high enough to support GAMM-produced smooth fits. Smooth terms appeared to be strongly influenced by small numbers of individuals in numerous interactions. It was therefore unclear whether the nonlinear relationships depicted, even when they appeared believable, could generalize to the population.

This issue was further complicated by the fact that conventional model-pruning methods, such as a comparison of models using the Akaike Information Criterion, are not a reliable technique to evaluate whether the degree of nonlinearity (i.e., "wiggliness") of the functions was suitable given the underlying data. These types of techniques attempt to arrive at the most parsimonious model by limiting the number of terms by the improvements in model likelihood that they provide (Akaike, 1974). However, a linear and nonlinear model built using GAMM could include an identical number of terms (where even several conjoined polynomial functions constitute a single smooth term), and the latter will always result in better likelihood as nonlinearities necessarily conform to individual variance (S. Wood, 2006). These methods are therefore not suitable for a comparison of the two, or even a comparison of nonlinear fits which vary in complexity.

Simulations of sample size variations suggested that doubling the number of participants would considerably alleviate over-fitting, at least in the context of the degree of variance found in the present data. In reality, should the within- and between-participant variance found in additional participants differ from our sample, this could influence the size of a sample that is considered adequate. This simulation can therefore only be used as a proof-of-concept that additional participants might help to alleviate concerns surrounding over-fitting. In addition, these findings suggest that an adequate sample size is also related to the complexity of the nonlinear relationship being modeled. Nonetheless, future research will need to be conducted in

this area to fully determine whether nonlinearities in these relationships are reliable, as well as meaningful.

## 7.4. Conditional Inference Random Forest Modeling

Our last investigation relied on a statistical method that represents a considerable departure from conventional regression techniques. This analysis used CForest, which combines elements of bootstrapping, permutation testing and hypothesis testing to deduce patterns in predictor variables which are systematically related to response variance (Breiman, 2001; Hothorn et al., 2006; Strasser & Weber, 1999; Strobl et al., 2009). CForest represents an amalgamation of techniques which attempt to address the limitations of additive modeling solutions in various ways. Most notably, the deductive permutation testing framework recursively partitions a series of randomized subsets of data into branching structures of isolated subsets. This branching of responses based on predictor variables provides a framework by which estimates can be assigned to novel observations, based on the predictor values associated with an observation, and repeating this process over hundreds of iterations results in a distribution of estimates around which confidence intervals can be computed. Importantly, this method does not assign estimates to fit terms to observations, and so all concerns surrounding the development of estimates or terms are irrelevant. Importantly, as estimates are calculated through averaging responses in individual branches, there is no assumption of linearity. In addition, the development of this branching structure is almost entirely data-driven except for a small number of user-defined constraints, which allows for definition of proficiency bins or electrode clusters.

A major aim of this analysis concerned the definition of regions in which the effect of IDs on response amplitude is depicted. The conventional approach to averaging electrodes over a

region involves deciding on a region or set of regions *a priori*, regardless of whether this grouping of electrodes will ideally capture the response's topography (e.g., Newman et al., 2012). Alternatively, electrode clustering solutions exist which define regions based on similarity of response amplitude, although this approach precludes describing gradual changes in topography which are associated with IDs (Pernet, Latinus, Nichols, & Rousselet, 2015b). Nonetheless, this type of clustering approach may be suited to certain types of research questions. Conversely, the recursive partitioning framework used by CForest allows for a data-driven delineation of electrode groupings, similarly based on response amplitude, but interactions with ID measures can be modeled and visualized. Using this method, the grouping of electrodes in which an effect should be depicted is deduced, rather than specified by the user, and can naturally follow the topographic distribution of an effect.

While this approach demonstrated advantages in groupings of categorical variables such as individual electrodes, it may not be ideally-suited to characterizing continuous variables (e.g., ID measures). The recursive partitioning mechanic is only capable of finding discrete divides in a variable in terms of how that variable predicts a response. For example, dividing individuals with scores above or below some metric into two subsets of data, based on this division maximizing the difference between the responses of those subsets. This type of division is naturally reflective of categorical variables, and can be used to define low- and high-proficiency groups of participants, but otherwise creates piecewise-constant stepping segments in continuous scales.

Notably, the response amplitude estimates in interactions which GAMM depicted using sinusoidal fits followed a similar fluctuating pattern, but the significance testing of these estimates (represented in the piecewise-constant segments overlaid in each figure) was largely either flat or included a single step where an overall trend existed. Therefore, this did appear to overcome the over-fitting issues that made interpretation of GAMM results difficult, and

therefore it may be suited to some research questions. However, we feel that the degree of characterization that either LME or GAMM could provide for continuous variables would be preferable for the typical ERP study. In particular, either regression approach seems better suited to describing the relationship between ID measures and the N400 or P600 amplitude. Ideally, the methods may be used in conjunction to define electrode groupings using CForest and the influence of ID measures in those regions using LME or GAMM. This combination of techniques will be described in greater detail below.

## 7.5. Interpreting Model Estimates and Significance

While the findings of the three approaches were relatively consistent in identifying the components of interest (i.e., the N400 and P600), there was some discrepancy in how each technique depicted the influences of ID measures. Where the presence and directionality of these effects have been discussed in previous chapters, overarching patterns in their sensitivity across the techniques will now be considered. Our initial analysis using LME will be used as a reference for models built using GAMM and CForest, as it is mechanically most similar to the analyses that are most common in this area (Liang and Chen, 2014; Moreno and Kutas, 2005; Newman et al., 2012; Pakulak and Neville, 2010; Tanner, 2013; Tanner et al., 2014; Tanner and Van Hell, 2014; Weber-Fox et al., 2003).

Overall, two characteristics were apparent in models built using GAMM when compared with LME. First, the violation effect (i.e., conditional contrast) amplitude was frequently lower in interactions between ID measures and the violation effect for any given ROI. Interestingly, GAMM resulted in model-predicted responses which fit the observed responses better did LME. While this may suggest that these lower-amplitude responses fit observed responses better, the improved model fit likely also resulted in part from nonlinearities conforming more closely to

individual responses. This pattern of conforming to individual responses is related to the second characteristic of interactions that were ubiquitous in these analyses, as response amplitudes for small numbers of individuals (often two or three) would dictate what appeared to be a strong nonlinear shift in a function which was otherwise unremarkable (i.e., there would otherwise be no overall trend aside from those individuals). As discussed, it is this pattern which raised concerns surrounding low sample sizes and over-fitting for this analysis.

Recall that for either LME or GAMM, an ID measure was concluded to have a significant effect if the conditional contrast term (LME) or either sentence type term (GAMM) was significant for an ROI, and the 95% CI of the contrast did not include 0 μV for at least one ID measure score. However, interpreting the significance of these interactions in GAMM models was further complicated by irregular confidence intervals, which frequently narrowed at the junction points of polynomial interactions (i.e., knots) and widened at extremes of an ID measure score. Given our current criteria for establishing significance, an ID measure could be concluded to significantly influence violation effect amplitude through weak fluctuations a mid-range scores, where confidence intervals were narrowest.

This interpretation can be problematic, as ID measure influences on violation effect amplitude that appeared relatively strong across the ID measure spectrum (3-5 μV), and which were significant when modeled using LME, were often non-significant in GAMM due to variability in confidence interval width. This pattern was evident, for example, in our GAMM model for phrase structure violations during 300-500 ms, specifically in the influence of Listening/Vocabulary scores on violation effect amplitude for the right anterior ROI. While no overall trend was evident, significant nonlinear smooth fits combined with narrow confidence intervals might suggest that the influence of this ID measure is significant on violation effect amplitude. As an alternative to our criteria selected for concluding significance of an ID

measure's influence on violation effect amplitude, future studies might consider either a higher-amplitude minimum response, or that the confidence intervals surrounding some proportion of participants all exceed 0 μV.

At times, interactions depicted using either LME or GAMM suggested entirely opposite influences of ID measures. For example, LME suggested that higher Speaking/Grammar scores were associated with a stronger P600 response to phrase structure violations. Conversely, GAMM suggested that Speaking/Grammar only influenced response amplitude in the lowest-scoring individuals, and that these individuals demonstrated a more negative violation effect than higher-scoring individuals. Our GAMM model depicted this effect was significant at several ROIs, but it was particularly strong in the anterior midline region. However, closer inspection revealed that this highly significant effect was driven largely by only a few participants, calling into question the validity of these findings. Similar patterns of conflicting responses were evident elsewhere, and as they were commonly driven by a small number of participants when using GAMM, they pattern may not replicate in a larger sample.

When interpreting three-way interactions between region, violation and an ID measure depicted using LME or GAMM, the reader must be mindful that these plots depicted partial effects. Partial effects in these 3x3 ROI plots depicted the overall violation effect for each ROI in addition to the effect of one ID measure, but do not include the partial effects of other ID measures. These interactions, as well as all random effects, summed to the total predicted response. That is, these partial effects contributed to the observed responses in a linearly additive manner. Therefore, if higher scores for two separate ID measures were each associated with higher response amplitude, this effect would have compounded in individuals who scored highly in both. However, the response amplitude depicted in each of these ID measure plots was only useful in interpreting the response amplitude increase that is associated with scoring highly

in one measure. In this way, attempts to interpret partial effects depicted by LME or GAMM might be misleading to the uninformed reader. Inclusion of the overall violation effect at each ROI ensures that interpreting these partial effects as significant (provided that 0 μV is not contained within the 95% CI) is valid, but only in interpretation of each ID measure's influence in isolation.

In this regard, CForest may be more easily interpretable than LME or GAMM estimates, where the response amplitude in an interaction reflects the raw response amplitude as it was delineated through permutation testing within subsets of the original data. That is, there are no partial effects, and as such effects need not sum to some overall estimate. The predicted response in an interaction modeled using CForest should ideally bear a one-to-one correspondence with recorded scalp voltage within any single branch of a tree. These responses were subsequently subjected to significance testing via CForest's CTree algorithm, which resulted in the characteristic stepping function that was overlaid on the forest estimates. That is, while considerable variability was seen in forest estimates, only variability in the stepping function was indicative of a significant trend. While there is a loss of information (i.e., variability across the ID measure spectrum) in this stepping function when compared with forest estimates, it maintains correspondence with observed scalp voltages, as there is no need for summation of partial effects in order to relate the two. With this in mind it is not surprising that CForest more frequently showed subsets of participants to show a weak violation effect, or none at all, where the summed partial effects of LME might also be insubstantial (or less substantial than might be implied by any single interaction term). Indeed, weaker response amplitude in interactions was frequently a characteristic of CForest's output when compared with LME or GAMM. This is also reflected in the overall violation effect topography of our LME analysis, which suggested that overall N400 and P600 response amplitudes were similar to those seen in the individual

interactions when using CForest (i.e., in the range of 2-5 μV). Therefore, the amplitudes of responses depicted in individual interactions using CForest more closely resembled the overall response amplitude, as the effects of each ID measure are not summed to provide an overall estimate, but instead reflect the mean response in any given subset of the full data set.

## 7.6. Implications for Future ERP Studies

These characteristics and patterns of sensitivity suggest that each of the above techniques might be best-suited for certain research questions, or that they might be used in conjunction to take advantage of the benefits that each technique provides. In particular, there are ways in which the characteristics of CForest might be leveraged to improve the sensitivity of regression models, and the two might be used in conjunction with relative ease. For example, where variables are categorical rather than continuous — such as individual electrodes or conditions or, in considering differences between individuals, if there were obvious groups such as a clinically-defined population and healthy controls — CForest is well suited to distinguish between categorical groupings in terms of how they might predict response amplitude. Describing the group of electrodes at which a response is significant represents an area where CForest might provide an improvement over traditional approaches (i.e., *a priori* ROI definition) with no apparent disadvantages. Not only are these electrode groupings free of restrictions in terms of size or distribution that might be imposed by an *a priori* definition, but their delineation is entirely data driven, resulting in findings that more accurately describe a particular data set. Furthermore, the user is still free to place any restrictions on size or distribution that might be desired.

Additional benefits of using CForest to describe differences in a response between levels of a categorical variable stem from its capabilities as a nonparametric testing procedure. For

example, CForest is less influenced by heteroscedasticity, as detailed in Chapter 5, and so it can more-accurately predict responses where the model has non-normally distributed residuals. This may occur if variance in response amplitude is inconsistent across ID measure scores, for example. As described in Chapter 3, while residuals in our LME models were not distributed such that they should render the model's findings invalid, neither were they perfectly normal. Therefore, CForest might be the most suitable approach to defining simpler categorical variables such as a violation effect, and in conjunction with electrode groupings might best describe the region over which a violation effect was elicited in a group of participants. Moreover, if the aim is to categorize participants into proficiency bins in some ID measures, then CForest is likely to be well-suited to this task as well.

Despite these advantages, two-way interactions between continuous variables such as ID measures and response amplitude lack the definition of regression models, and thus there is a loss of both predictive and descriptive power on these axes. This is because the binary splits that CForest performs result in piecewise-constant predictions for continuous variables, rather than a linear slope or smooth fit of conjoined polynomial functions that regression techniques can provide. Therefore, either LME or GAMM might be better-suited to describing the influence of ID measures on violation effect amplitude. However, the two need not be mutually exclusive. Indeed, future studies might use a combination of approaches, whereby the two-way interaction between violation effect and region is defined using CForest in order to narrow the scope of findings to an electrode grouping at which a component is identified. Following this, a regression model might be used to describe the influence of ID measures on response amplitude for those electrodes at which the violation effect is significant. This type of analysis would benefit from all of the advantages of CForest in detecting data-driven regions of interest which better fit the

topography of an effect, but also maintain the more precise descriptive power of continuous ID measure influence that regression can provide.

When describing the influence of ID measures, the choice of whether to allow for nonlinearities in these effects should be based on sample size and distribution of ID measure scores. Our simulations in Chapter 4 have suggested that doubling the present sample size to include approximately 60 participants might be appropriate given the distribution of ID measures and responses that were observed. It is important to bear in mind that this number cannot be used as a general rule and the propensity for GAMM to over-fit to observations will depend on specific characteristics of a data set's distribution. In either case, combining the CForest with regression models in this way may provide a powerful framework for describing the influence of ID measures in electrode groupings which are specific to the response being investigated.

In cases such as the present data set, with all considerations specific to our sample size, ID measures used and distribution of data (both responses and ID measures) the ideal analytical procedure would likely be one which combines the data-driven ROI selection of CForest (using a model which includes only the two-way condition by electrode interaction) with modeling ID measure influence within the resulting electrode grouping using LME. For this approach, our AIC-driven model selection framework represents an adequate means of circumventing concerns surrounding multicollinearity, with the unfortunate disadvantage that some ID measures (in our case, typically AzBio, TOWRE, and LSpan) are necessarily excluded from analysis. However, should these be measures of interest, they could still be included, provided that their collinear counterparts are not.

Our model specification supporting a maximum of three-way interactions between condition, ROI, and each ID measure is likely sufficient to describe the influence of ID measures

on violation effect amplitude at each ROI, and we suspect that describing interactions between ID measures is not likely to be of interest in many areas. Therefore, higher-level interactions should not be necessary. However, limiting this analysis to a single ROI as determined by CForest means that these three-way interactions could be reduced to two-way interactions between condition and each ID measure of interest, where all electrodes within the ROI are considered. The result would be estimates which are specific to the region where a violation effect is strongest, and with no need for multiple comparison correction across ROIs.

Lastly, our investigations into random effect structure suggested that allowing for random variability in scalp topography across participants, as well as random variability in mean response across participants, produced a considerable improvement in model likelihood over simpler random effect structures, or none at all. However, again with topography no longer a consideration in model terms, a simplified random effect structure that includes only random participant means should suffice. Given that the improvement which resulted from including a random effect of scalp topography within participants was found in a model which included the full scalp, limiting the scope of data to only the ROI of interest may in part circumvent the issue of scalp topography variability between participants.

## 7.7. Suitability of ID Measures

The ID measures that were selected for investigation in the present analyses were chosen in part because they have been associated with N400 and P600 amplitude in the past (Newman et al., 2012; Pakulak & Neville, 2010), or because they were theorized to index cognitive functions which have been related to these components, in the case of working memory (Nakano et al., 2010). In the case of our AzBio speech comprehension task and the TOWRE word reading efficiency task, the outcomes of these measures were unknown, and their

inclusion was largely exploratory. These analyses have provided additional support for the link between ID measures and the amplitude and topography of language-related ERP components. However, the tasks that were selected may not have been ideal in terms of maximally characterizing differences between individuals. In the case of the TOAL-3, for example, which provided many of the ID measures which were found to significantly influence N400 and P600 amplitude (Speaking/Grammar, Listening/Grammar, and Listening/Vocabulary) this test was designed to identify areas of relative strength and weakness, particularly in those who might benefit from programs for language intervention (Hammill et al., 1994). These tasks were therefore intended to characterize broad strokes of language proficiency across several areas for practical learning purposes, but not necessarily to capture nuanced multidimensional descriptions of grammatical ability.

Nonetheless, the TOAL-3 subtests provided the most consistently significant predictors of response amplitude across the sentence types and time windows that were investigated, and so broad strokes of language proficiency in this case may be suitable to detect associations with cortical processing. The possibility remains, however, that more detailed descriptions of proficiency could further separate individuals on dimensions that are meaningfully related to cortical processing. Other ID measures, such as AzBio, were only very rarely identified as significant predictors of response amplitude. Recall that our distribution of AzBio scores was strongly skewed toward the high end with low variance, and only included a small number of lower-scoring individuals. In this case, the task difficulty may have been too low to reliably detect differences between individuals in terms of speech comprehension. The fact that this ID measure was found to at least weakly predict response amplitude is encouraging, however, and perhaps alternative measures of speech comprehension should be considered and investigated in future research.

Beyond the question of selecting ID measures which could maximize nuanced differences between individuals in terms of language proficiency or other aspects of cognition, there is also the question of which cognitive constructs are being indexed by these measures. There is a degree of overlap in the measures selected, as made evident by the multicollinearity outlined in previous chapters. For example, one might rationally expect the parsing of complex sentences into understandable units – as indexed through our assessments of grammatical ability – to rely in part on working memory capacity. It is no surprise then that the highest degree of multicollinearity seen was between our TOAL-3 Listening/Grammar subtest and our Listening Span measure of working memory capacity. However, despite similarities these measures may have in predicting task performance, or in their correlations across participants, they have demonstrated a high degree of validity and/or reliability in the specific avenues of cognitive ability that they are intended to address. Specifically, the TOAL-3 subtests which were used in each of our analyses have been demonstrated to have strong content validity in their respective domains (Hammill et al., 1994). Similarly, the OSpan task was central to our hypotheses regarding working memory capacity, and has demonstrated strong test-retest reliability (Unsworth et al., 2005). These tasks were therefore considered appropriate for our aims of relating the aspects of cognition which they were designed to index with performance on two specific types of violation assessments, and corresponding brain activity.

## 7.8. Limitations of This Research

While the present work aimed to provide a variety of methodological improvements to ERP studies concerned with ID measures, it also made apparent some limitations and considerations for which there are no apparent solutions at this time. First, as discussed, continuous variables might be best-described through regression approaches. However, by their

nature, regression models work best with orthogonal sets of predictors, and so some measure to circumvent multicollinearity must be used. This might include removal of ID measures which are too similar to others but yield less viable models, which was the approach used here. Alternatively, collinear predictors might be combined into a single measure, through averaging or some other means. In either case the result is a loss of the ability to assign meaningful differences in responses to specific ID measures, either because they have been removed, or because they have been combined irreversibly with others. It is therefore very possible that predictors which have been excluded might be significantly related to responses, but this cannot be known. Using the present approach, they could only be included at the expense of those predictors which were shown to contribute more strongly to the model's log-likelihood. While multicollinearity is considerably less harmful to a model's predictive ability when using CForest, as discussed in Chapter 5, the same approach toward predictor exclusion was used in all methods to maximize our ability to compare results between the three techniques.

One additional limitation is that, when presented with evidence from multiple analytical approaches that suggest different – sometimes mutually exclusive – findings, it is not clear which may be reliable. In some cases, this was the result of our model selection process. For example, eliminating collinear predictor variables which contributed less strongly to the model precluded finding and association between those predictors and violation effect amplitude. However, as this was not required for analysis with CForest, eliminated predictors occasionally surfaced as having significant influence on responses. In other cases, however, the directionality of influence was entirely reversed between our LME and GAMM models. While we suspect that this may have stemmed from smooth fits which conformed strongly to individual variance, critically, whether this is indeed the case cannot be known. Further investigation in larger samples will be required to elucidate the true nature of these relationships.

It is important to be mindful of test-retest reliability in any ERP study. While it has been shown that individuals generally perform consistently in EEG studies of cognition, and that ERP responses to a working memory task collected seven days apart have shown a strong correlation (r=0.90), some degree of variability is still expected (McEvoy, Smith, & Gevins, 2000). In particular, ERP responses are sensitive to changes in the cognitive state of individuals, which may be related to diet, sleep patterns, or other factors. This highlights the need for replication of findings, both within and between populations. Moreover, efforts should be made to test individuals who are representative of the populations to which findings are generalized, given that these factors may vary systematically with population characteristics (e.g., age). Therefore, given the limited representation of individuals in the present research, it cannot be certain that our findings will hold in the general population.

A related issue is that, while individual differences were of central interest to the present set of investigations, identifying these differences may have been limited by the number of participants included in our models. Beyond filling gaps in the distribution of ID measure scores, a number of scores were frequently seen in single participants, which provided considerable weight to the fit of estimates in association with specific scores. This likely contributed to instances of over-fitting. In this regard, a larger sample of participants would likely benefit our findings.

In addition to sample size, the above concerns surrounding over-fitting may have resulted from improper random effect specification. Specifically, while observations were modeled for individual trials, a random effect of trial was not included in our models. This may have resulted in two outcomes in our results. First, without a random effect accounting for individual trials, it is possible that error variance was not equally and independently distributed across our model terms, resulting in errors in estimation of *p*-values. This problem may have

impacted our findings for both LME and GAMM, though our inability to report the significance of difference curves using GAMM meant that this limitation would not have hindered our interpretation of smooth fit significance in this regard. Second, failure to include a random effect of trial may have contributed to over-fitting, as trials were otherwise treated as repeated measures within participants, placing considerably more weight on predictors associated with individual participants (e.g., specific ID measure scores). Modeling a random effect of trial might be expected to reduce the tendency for smooth fits to conform to the responses of each individual, while simultaneously narrowing confidence intervals, as error variance is reduced. As an alternative to including a random effect of trial, averaging responses over trials for each condition in each sentence type and time window might also mitigate issues surrounding over-fitting, but would result in a considerably reduced data set and a corresponding loss of statistical power. Failure to account for inter-trial variability may have impacted findings both for LME and GAMM, but given the flexibility of smooth fits using GAMM, it is more likely to have affected outcomes in the latter. The result in both cases may have been an increase in Type I error. With these considerations in mind, the conclusions drawn regarding the use of GAMM in ERP data may be specific to the present data set.

Lastly, due to the concerns surrounding sample size as well as random effect specification, it is still unclear whether relaxing the assumption of linearity could elucidate meaningful details regarding the relationships between ID measures and the violation effect response. Our findings suggested that variance in the N400 or P600 response amplitude might be limited to portions of ID measure spectrums, but a number of associations appeared erroneous (i.e., primarily those which were depicted as sinusoidal, but potentially even those findings which appeared to conform to expectations). These inconsistencies shrouded doubt on the efficacy of findings that stemmed from GAMM. However, this may remain a fruitful area for

future research, and a more robust description of nonlinearities may inform future attempts to model the link between ID measures and ERPs.

## 7.9. Future Developments

The suggested combination of approaches provides a number of advantages associated with each technique, while aiming to minimize the disadvantages that each is limited by. That said, some difficulties will still need to be overcome before an ideal modeling process is available. First, while CForest is capable of including an entire set of ID measures, despite multicollinearity, this characteristic cannot be taken advantage of if ID measures are described using regression. While we feel that the model selection process defined in Chapter 3 provides an adequate framework for handling multicollinearity, it necessarily excludes some measures. Therefore, the ideal analysis might describe ID measures with the nuance of regression while avoiding a framework which requires terms to compete for the definition of variance (i.e., additive modeling). This might be achieved through variations in how significance of a forest's distribution is determined. At present, the algorithm which produces individual trees is used with the forest's distribution of estimates to determine significance. However, this is what produces the stepping functions associated with CForest, and so alternative approaches might be explored. While development of new statistical procedures is beyond the scope of this dissertation, smooth forest estimates from CForest would represent a powerful advantage in descriptive analytics.

While the present study evaluated response amplitude averaged within two time windows of interest, treating time as a continuous variable may yield some benefit. For example, this might allow the user to detect whether differences in ID measure scores are related to differences in onset latency for a response. Indeed, it has been suggested that higher-proficiency

individuals show an earlier P600 onset in response to phrase structure violations (Pakulak & Neville, 2010). This finding was achieved through comparing lower- with higher-proficiency individuals in two bins, rather than treating proficiency as a continuous variable. Assessing the interaction between ID measures and time, where each is continuous, as well as condition and region, might yield interesting and meaningful interactions. Indeed, aside from the above limitations of CForest's descriptive abilities, it might present an adequate means to evaluate significance in such interactions.

Lastly, the approaches taken in these analyses are best-suited to an exploratory investigation in which little is known about the underlying truth of a relationship between an ID measure and response – be it in topography, latency, amplitude, or the shape of the function that relates the two. Therefore, at each step numerous models were fitted with the goal of uncovering the ideal modeling approach for any effects that become apparent. However, this is counter to a confirmatory approach in which parameters should ideally specified at the outset and sensitivity to the effects of interest relies on selecting time windows, regions and modeling parameters using *a priori* knowledge. The aspect of our LME and GAMM analyses which would be most problematic in this regard was our framework for addressing multicollinearity, as this relied on iteratively producing increasingly-complex models until no further predictors could justifiably be added. Instead, a confirmatory analysis might opt to restrict predictors of interest to those which pertain to the specific research question, or to use alternative approaches to model selection such as elastic net regularization (Zou & Hastie, 2005). Moreover, while our central hypotheses pertained to the earlier time window (300-500 ms) during presentation of semantic violations, and the later time window (600-800 ms) during presentation of phrase structure violations, all four combinations of time windows and sentence types were

investigated for the sake of completeness. Again, such an exhaustive approach may not be required in the context of a confirmatory analysis.

## 7.10. Conclusions

The present research aimed to evaluate the applicability of several statistical techniques to describing the relationships between a number of ID measures and ERP component characteristics. This includes attempts to optimize the model building and selection process in language violation processing through a number of means. Findings have demonstrated that inclusion of random effects can result in considerable improvement to model fit when compared with fixed-effects-only methods that have been used in prior research, and have also provided a framework for addressing multicollinearity among ID measures. While results suggested that model sensitivity might be further improved through relaxing the assumption of linearity in the effect of ID measures on response amplitude, these findings will need to be validated in data sets with a larger sample size. Moreover, CForest provided a means to perform significance testing of a response's topography, both in terms of size and distribution, even among interactions with specific ID measures. Beyond simply optimizing user-specified aspects of a model, these latter analytical improvements represent a shift from *a priori* specification of hypotheses to more data-driven effect descriptions. Such a change in analytical approach might be used to explore not only an effect's distribution in space (i.e., across the scalp), as in the present research, but future studies might also explore the utility of these techniques in evaluating response latency as a continuous variable.

# References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. (2015). Parsimonious mixed models. *arXiv:1506.04967*.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes. Retrieved from http://cran.r-project.org/package=lme4

Belsley, D. A. (1984). Demeaning Conditioning Diagnostics through Centering. *The American Statistician*, *38*(2), 73. https://doi.org/10.2307/2683236

Bollinger, G., Belsley, D. A., Kuh, E., & Welsch, R. E. (1981). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *Journal of Marketing Research*, *18*(3), 392. https://doi.org/10.2307/3150985

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bry, X., Trottier, C., Mortier, F., Cornu, G., & Verron, T. (2016). Supervised Component Generalized Linear Regression with Multiple Explanatory Blocks: THEME-SCGLR (pp. 141–154). https://doi.org/10.1007/978-3-319-40643-5_11

Bryll, A., Binder, M., & Urbanik, A. (2013). The influence of proficiency level of foreign language on the activation patterns of language areas. *Przegląd Lekarski*, *70*(5), 243–247. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23944090

Burnham, K. P., & Anderson, D. R. (2004). *Model Selection and Multimodel Inference*. (K. P. Burnham & D. R. Anderson, Eds.). New York, NY: Springer New York. https://doi.org/10.1007/b97636

Cardinali, C. (2013). Data Assimilation: Observation influence diagnostic of a data assimilation system. Shinfield Park, England: European Centre for Medium-Range Weather Forecasts: Research Department.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and Brain Sciences*, *31*(5), 489-508-58. https://doi.org/10.1017/S0140525X08004998

Cobb, W. A., & Dawson, G. D. (1960). The latency and form in man of the occipital potentials evoked by bright flashes. *Journal of Physiology*, *152*(1), 108–121.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15. https://doi.org/10.2307/1268249

Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition*, *30*(6), 958–968. https://doi.org/10.3758/BF03195780

Dalal, D. K., & Zickar, M. J. (2012). Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods*, *15*(3), 339–362. https://doi.org/10.1177/1094428111430540

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics* (1st editio). New York: Oxford University Press.

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

Dong, S., Reder, L. M., Yao, Y., Liu, Y., & Chen, F. (2015). Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. *Brain Research*, *1616*, 146–156. https://doi.org/10.1016/j.brainres.2015.05.003

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, *16*, 143–149.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00335

Felder, R. M., & Soloman, B. A. (n.d.). Index of Learning Styles.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6356204

Fodor, J. (1983). *Modularity of the Mind*. Cambridge, MA, MA: MIT Press.

Grandvalet, Y. (2004). Bagging Equalizes Influence. *Machine Learning*, *55*(3), 251–270. https://doi.org/10.1023/B:MACH.0000027783.34431.42

Green, B. F., & Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika*, *25*(2), 127–152. https://doi.org/10.1007/BF02288577

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Hammill, D., Brown, L., Larsen, S., & Wiederholt, L. (1994). *Test of Adolescent Language, third edition (TOAL-3)*. Austin: PRO-ED.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hillyard, S. A., & Munte, T. F. (1984). Selective attention to color and location: An analysis with event-related brain potentials. *Perception & Psychophysics*, *36*(2), 185–198. https://doi.org/http://dx.doi.org/10.3758/BF03202679

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441. https://doi.org/10.1037/h0071325

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York, NY, NY: Springer New York. https://doi.org/10.1007/978-0-387-47946-0

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2000). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *111*(10), 1745–1758. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11018488

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

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122–149. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1546114

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: an electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean Centering in Moderated Multiple Regression: Much Ado about Nothing. *Educational and Psychological Measurement*, *58*(1), 42–67. https://doi.org/10.1177/0013164498058001005

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

Kuperberg, G. R., 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. https://doi.org/10.1016/j.bandl.2005.12.006

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

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–208.

Kutas, M., & Van Petten, C. (1994). ERP Psycholinguistics electrified: Event-related brain potential investigations. In *Handbook of psycholinguistics* (pp. 83–143). San Diego: Academic Press. https://doi.org/10.1016/B978-012369374-7/50018-3

LaMotte, L. R. (2014). Fixed-, Random-, and Mixed-Effects Models. In *Wiley StatsRef: Statistics Reference Online*. Chichester, UK, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat03169

Lefly, D. L., & Pennington, B. F. (2000). Reliability and validity of the adult reading history questionnaire. *Journal of Learning Disabilities*, *33*(3), 286–296. https://doi.org/10.1177/002221940003300306

Li, Y., Peng, D., Liu, L., Booth, J. R., & Ding, G. (2014). Brain activation during phonological and semantic processing of Chinese characters in deaf signers. *Frontiers in Human Neuroscience*, *8*(April), 211. https://doi.org/10.3389/fnhum.2014.00211

Liang, L., & Chen, B. (2014). Processing morphologically complex words in second-language learners: the effect of proficiency. *Acta Psychologica*, *150*, 69–79. https://doi.org/10.1016/j.actpsy.2014.04.009

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews. Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

LimeSurvey-Development-Team. (2012). LimeSurvey - The free and open source survey software tool. Retrieved from http://www.limesurvey.org/

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 381–400. https://doi.org/10.1111/1467-9868.00183

Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. (1996). Independent component analysis of Electroencephalographic data. *Advances in Neural Information Processing Systems*, *8*, 145–151.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*(1), 4–18. https://doi.org/doi.org/10.1111/j.1469-8986.1995.tb03400.x

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*, 940–967. https://doi.org/1092-4388/07/5004-0940

McEvoy, L. K., Smith, M. E., & Gevins, A. (2000). Test-retest reliability of cognitive EEG. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *111*(3), 457–463. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10699407

McWhinney, S. R. S. R., Tremblay, A., Chevalier, T. M. T. M., Lim, V. K. V. K., & Newman, A. J. A. J. (2016). Using CForest to Analyze Diffusion Tensor Imaging Data: A Study of White Matter Integrity in Healthy Aging. *Brain Connectivity*, *6*(10), 747–758. https://doi.org/10.1089/brain.2016.0451

Meyer, A., Lerner, M. D., De Los Reyes, A., Laird, R. D., & Hajcak, G. (2017). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology*, *54*(1), 114–122. https://doi.org/10.1111/psyp.12664

Moreno, E. M., & Kutas, M. (2005). Processing semantic anomalies in two languages: an electrophysiological exploration in both languages of Spanish-English bilinguals. *Brain Research. Cognitive Brain Research*, *22*(2), 205–220. https://doi.org/10.1016/j.cogbrainres.2004.08.010

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R 2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nakano, H., Saron, C., & Swaab, T. Y. (2010). Speech and span: working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience*, *22*(12), 2886–2898. https://doi.org/10.1162/jocn.2009.21400

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience*, *24*(5), 1205–1223. https://doi.org/10.1162/jocn_a_00143

Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An ERP study of regular and irregular English past tense inflection. *NeuroImage*, *34*(1), 435–445. https://doi.org/10.1016/j.neuroimage.2006.09.007

Newman, S. D., Malaia, E., Seo, R., & Cheng, H. (2013). The Effect of Individual Differences in Working Memory Capacity on Sentence Comprehension: An fMRI Study. *Brain Topography*, *26*(3), 458–467. https://doi.org/10.1007/s10548-012-0264-8

Nobre, A. C., & McCarthy, G. (1994). Language-Related ERPs: Scalp Distributions and Modulation by Word Type and Semantic Priming. *Journal of Cognitive Neuroscience*, *6*(3), 233–255. https://doi.org/10.1162/jocn.1994.6.3.233

Ojima, S., Nakata, H., & Kakigi, R. (2005). An ERP Study of Second Language Learning after Childhood: Effects of Proficiency. *Journal of Cognitive Neuroscience*, *17*(8), 1212–1228. https://doi.org/10.1162/0898929055002436

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5146491

Osterhout, L. (1997). On the Brain Response to Syntactic Anomalies: Manipulations of Word Position and Word Class Reveal Individual Differences. *Brain and Language*, *59*(3), 494–522. https://doi.org/10.1006/brln.1997.1793

Osterhout, L., & Mobley, L. (1995). Event-Related Brain Potentials Elicited by Failure to Agree. *Journal of Memory and Language*, *34*(6), 739–773. https://doi.org/10.1006/jmla.1995.1033

Pakulak, E., & Neville, H. J. (2010). Proficiency Differences in Syntactic Processing of Monolingual Native Speakers Indexed by Event-related Potentials. *Journal of Cognitive Neuroscience*, *22*(12), 2728–2744. https://doi.org/10.1162/jocn.2009.21393

Parvinnia, E., Sabeti, M., Zolghadri Jahromi, M., & Boostani, R. (2014). Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *Journal of King Saud University - Computer and Information Sciences*, *26*(1), 1–6. https://doi.org/10.1016/j.jksuci.2013.01.001

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015a). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015b). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Rashid, N. A., Taib, M. N., Lias, S., Sulaiman, N., Murat, Z. H., & Kadir, R. S. S. A. (2011). Learners' Learning Style Classification related to IQ and Stress based on EEG. *Procedia - Social and Behavioral Sciences*, *29*, 1061–1070. https://doi.org/10.1016/j.sbspro.2011.11.339

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, *6*(1), 15–32.

Sánchez, J. (1982). Multivariate Analysis. *Biometrical Journal*, *24*(5), 502–502. https://doi.org/10.1002/bimj.4710240520

Simon, J. R., & Barbaum, K. (1990). Effect of conflicting cues on information processing: The "Stroop effect" vs. the "Simon effect." *Acta psychologica1*, *73*(159–170). https://doi.org/10.1016/0001-6918(90)90077

Spahr, A. J., Dorman, M. F., Litvak, L. M., Cook, S. J., Loiselle, L. M., DeJong, M. D., … Gifford, R. H. (2014). Development and Validation of the Pediatric AzBio Sentence Lists. *Ear and Hearing*, *35*(4), 418–422. https://doi.org/10.1097/AUD.0000000000000031

Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: evidence from event-related brain potentials. *Second Language Research*, *25*(1), 13–41. https://doi.org/10.1177/0267658308098995

Strasser, H., & Weber, C. (1999). On the asymptotic theory of permutation statistics, *8*, 220–250.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Tanner, D. (2013). Individual differences and streams of processing. *Linguistic Approaches to Bilingualism*, *3*(3), 350–356. https://doi.org/10.1075/lab.3.3.14tan

Tanner, D., Inoue, K., & Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism: Language and Cognition*, *17*(2), 277–293. https://doi.org/10.1017/S1366728913000370

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, *56*, 289–301. https://doi.org/10.1016/j.neuropsychologia.2014.02.002

Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin: PRO-ED.

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, *52*(1), 124–139. https://doi.org/10.1111/psyp.12299

Tremblay, A., & Ransijn, J. (2013). LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions.

Turano, M. T., Marzi, T., & Viggiano, M. P. (2016). Individual differences in face processing captured by ERPs. *International Journal of Psychophysiology*, *101*, 1–8. https://doi.org/10.1016/j.ijpsycho.2015.12.009

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16405146

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

van Rij, J., Wieling, M., Bayyen, R., & van Rijn, H. (2017). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.

Van Voorhis, S., & Hillyard, S. A. (1977). Visual evoked potentials and selective attention to points in space. *Perception & Psychophysics*, *22*(1), 54–62.

Vos, S. H., Gunter, T. C., Kolk, H. H., & Mulder, G. (2001). Working memory constraints on syntactic processing: an electrophysiological investigation. *Psychophysiology*, *38*(1), 41–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11321620

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & review2*, *11*(1), 192–196.

Weber-Fox, C., Davis, L. J., & Cuadrado, E. (2003). Event-related brain potential markers of high-language proficiency in adults. *Brain and Language*, *85*(2), 231–244.

Wood, S. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(3), 495–518. https://doi.org/10.1111/j.1467-9868.2007.00646.x

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). CRC/Taylor & Trancis.

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistics Society*, *64*(1), 139–155.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Wood, S. N. S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 413–428. https://doi.org/10.1111/1467-9868.00240

Zainuddin, Z., Huong, L. K., & Pauline, O. (2012). On the Use of Wavelet Neural Networks in the Task of Epileptic Seizure Detection from Electroencephalography Signals. *Procedia Computer Science*, *11*, 149–159. https://doi.org/10.1016/j.procs.2012.09.016

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x